Journal of Multi Disciplinary Engineering Technologies
Volume 19, Issue 01
DOI: https://doi.org/10.64992/jmdet.vol.19.issue.1.4

Deep Learning Approaches for Multimodal Emotion Recognition: Trends, Issues, and Prospects

Kirti Sharma[1*] and Rainu Nandal[2]

[1]University Institute of Engineering Technology, Maharshi Dayanand University, Rohtak, Haryana, India.

*Corresponding author(s). E-mail(s): krtbhardwaj1@gmail.com;
Contributing authors: rainunandal.uiet@mdurohtak.ac.in;

Abstract

High-fidelity human-computer interaction now hinges on the machine's ability to decode affective states—a task where single-source data (speech or text alone) consistently falls short. This paper pivots away from traditional unimodal constraints, centering instead on Multimodal Emotion Recognition (MER) as the primary vehicle for context-aware intelligence. We provide a rigorous deconstruction of current deep learning frameworks, specifically interrogating how different fusion topologies and network architectures handle cross-modal interference. Beyond a simple performance review, this survey exposes the "cracks" in state-of-the-art systems: the persistent struggle with cultural bias, the scarcity of high-quality labels, and the inherent opacity of deep-stack models. Our findings suggest that the next frontier lies not in larger models, but in "low-rank" self-supervised learning and Explainable AI (XAI). By prioritizing these lean, transparent methodologies, the field can finally move toward emotion-aware tech that is both ethically robust and deployable on the edge.

Keywords: Multimodal Emotion Recognition, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks, Transformers, Fusion, Affective Computing

1 Introduction

Human interaction is fundamentally driven by emotional intelligence, influencing everything from high-level decision-making to basic social cohesion. For artificial intelligence to truly integrate into the human experience—whether through assistive healthcare, adaptive pedagogy, or social robotics—it must move beyond simple data processing and begin to parse affective nuances [1, 2]. Historically, emotion detection relied on unimodal frameworks. However, isolating facial geometry, vocal prosody, or lexical sentiment often yields a fragmented and brittle understanding of human intent. A smile may be deceptive, a tone ambiguous, and text famously fails to capture sarcasm [2, 3]. Because these single-channel systems cannot resolve such contradictions, they frequently collapse when exposed to the messiness of real-world environments [1, 4]. Multimodal Emotion Recognition

Published on: Dec, 2025 ISSN(P) 0974-1771 ISSN(O) 2581-9372

(MER) offers a solution by synthesizing divergent streams—audio, visual, textual, and even physiological—into a cohesive interpretive framework [1]. By cross-referencing these heterogeneous signals, MER can "fill the gaps" left by any single modality, drastically boosting both system reliability and predictive accuracy [5]. The surge in deep learning has effectively weaponized MER capabilities. While CNNs handle spatial imagery and LSTMs manage the temporal flow of speech, the recent emergence of Transformer-based attention mechanisms has redefined how we model cross-modal interplay [4, 6]. These architectures facilitate end-to-end pipelines that bypass the need for tedious, manual feature extraction, instead distilling complex emotional representations directly from the raw data [5]. Yet, the field has not reached its zenith. Significant friction remains: high-quality, culturally diverse data is notoriously scarce, and models often struggle with "noisy" inputs or missing modalities in real-time scenarios. Furthermore, the "black box" nature of these neural networks complicates their use in sensitive sectors where transparency is non-negotiable. This paper provides a comprehensive review of deep learning approaches in MER, focusing on: 1. Benchmark datasets and their characteristics. 2. Network architectures and fusion strategies. 3. Comparative analysis of state-of-the-art methods. 4. Key challenges and emerging research directions. By synthesizing progress and identifying open issues, this review aims to guide researchers and practitioners toward developing more robust, interpretable, and culturally inclusive MER systems.

2 The Modalities and Representations

Multimodal Emotion Recognition (MER) is based on the integration of multiple information sources, as each modality captures different aspects of human affective states. The primary modalities considered in MER studies include visual, audio, and textual signals, which together provide complementary and enriched emotional information.

2.1 Visual Modality

The visual modality primarily focuses on facial expressions, head movements, eye gaze, and body gestures. Facial Action Units (FAUs) and micro-expressions offer subtle indicators of emotional states, while full-body gestures can convey additional contextual information [7]. Deep learning architectures, particularly Convolutional Neural Networks (CNNs) and their variants such as ResNet and VGG, have demonstrated high effectiveness in extracting spatial features from both static images and video sequences for emotion recognition tasks [1, 8].

2.2 Audio Modality

Emotional information in speech is conveyed through variations in prosody, pitch, loudness, and spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) [2]. Temporal modeling of these acoustic features is efficiently handled by Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures, which capture sequential dependencies in audio signals [4, 6]. Recent deep learning approaches enable end-to-end processing of raw audio data, thereby reducing reliance on manually engineered features.

2.3 Text Modality

Text-based emotion recognition relies on semantic meaning, syntactic structure, and sentiment expressed through language [3]. Transformer architectures and pre-trained language models, particularly BERT, have emerged as powerful tools for textual emotion analysis. These models are capable of capturing complex linguistic phenomena such as sarcasm, implicit emotions, and contextual dependencies that are often difficult to infer from visual or auditory cues alone [5].

2.4 Complementarity of Modalities

Multimodal systems integrate multiple input channels to exploit their complementary strengths, thereby improving prediction accuracy and robustness. For instance, when facial expressions are occluded or verbal expressions are ambiguous, textual information may serve as an alternative emotional cue. Effective fusion of heterogeneous modalities requires precise temporal and spatial synchronization, which is commonly achieved using deep learning techniques such as attention mechanisms and transformer-based fusion networks.

3 Benchmark Datasets

Progress in multimodal emotion recognition relies heavily on benchmark datasets that provide standardized resources for training, evaluation, and fair comparison of models. Most MER datasets include audio, visual, and textual modalities, while some also incorporate physiological signals. A summary of commonly used benchmark datasets is presented in Table 1.

3.1 Dataset Characteristics

- **IEMOCAP**: A benchmark dataset for speech and visual-based MER that combines scripted dialogues with spontaneous interactions, enabling models to learn naturalistic emotional expressions.
- **RAVDESS**: Provides high-quality audiovisual recordings collected under controlled conditions and is commonly used to evaluate model robustness on acted emotional expressions.
- **CMU-MOSEI**: A large-scale multimodal sentiment dataset containing text, audio, and visual information. It is widely used for training and evaluating transformer-based MER models.
- **MELD**: Focuses on emotion recognition in multi-party conversational settings, capturing the inherent complexity of real-world dialogues.
- **MSP-IMPROV**: Combines improvised and acted speech, making it particularly suitable for modeling temporal dynamics of emotional expressions.
- **SEMAINE**: Consists of naturalistic interactions between humans and virtual agents, supporting the study of continuous and perceptual affective states.

3.2 Observations and Challenges

- Most publicly available MER datasets are English-centric, limiting their ability to generalize across cultures and languages.
- A significant number of datasets rely on acted emotions rather than spontaneous expressions, which may introduce bias in model performance.
- Accurate temporal alignment across modalities is crucial for MER; however, fine-grained synchronization is not consistently available across datasets.
- Physiological modalities remain underrepresented, highlighting an opportunity for more comprehensive multimodal emotion analysis.

These datasets form the foundation for training, evaluating, and benchmarking deep learning-based MER models. Large-scale multimodal corpora such as CMU-MOSEI and MELD are increasingly used to validate transformer-based and attention-driven architectures.

4 Multimodal Emotion Recognition (MER) Fusion Strategies

The objective of Multimodal Emotion Recognition (MER) is to infer an individual's emotional state using heterogeneous sources such as facial expressions, speech, textual content, and physiological signals. The strategy used to integrate these modalities, known as the fusion strategy, plays a critical

Table 1: Benchmark Datasets Description

Dataset	Size / Citation	Details
IEMOCAP	12 hours; Busso et al., 2008	**Modalities:** Audio, Visual, Text **Emotion Labels:** 9 emotions (anger, happiness, sadness, neutral, etc.) **Remarks:** Conversational dataset with scripted and spontaneous recordings
RAVDESS	7,356 clips; Livingstone & Russo, 2018	**Modalities:** Audio, Visual **Emotion Labels:** 8 emotions (neutral, calm, happy, sad, angry, fearful, surprise, disgust) **Remarks:** High-quality acted audiovisual emotional speech and song recordings
CMU-MOSEI	23,453 clips; Zadeh et al., 2018	**Modalities:** Audio, Visual, Text **Emotion Labels:** 6 sentiment labels (positive, negative, neutral) **Remarks:** Large-scale dataset with diverse speakers, topics, and expressive styles
MELD	13,000 utterances; Poria et al., 2019	**Modalities:** Audio, Visual, Text **Emotion Labels:** 7 emotion categories **Remarks:** Dialogue-based dataset with multi-speaker, multi-turn conversations
MSP-IMPROV	8,438 clips; Busso et al., 2017	**Modalities:** Audio, Visual **Emotion Labels:** 4 emotions (happy, sad, angry, neutral) **Remarks:** Semi-naturalistic recordings with varied and improvised acting styles
SEMAINE	959 conversations; McKeown et al., 2012	**Modalities:** Audio, Visual **Emotion Labels:** 5 affective states **Remarks:** Naturalistic human–agent interaction scenarios

role in leveraging complementary information, suppressing noise, and improving model robustness. Existing fusion approaches can be broadly categorized into feature-level (early) fusion, decision-level (late) fusion, hybrid fusion, and attention/transformer-based fusion.

4.1 Feature-Level (Early) Fusion

In feature-level fusion, features extracted from each modality are combined into a unified representation prior to classification. Let X_v, X_a, and X_t represent visual, audio, and textual feature vectors, respectively. Feature-level fusion can be expressed as:

$$X_f = f(X_v, X_a, X_t) \tag{1}$$

where $f(\cdot)$ denotes a fusion function such as concatenation, bilinear pooling, or projection into a shared embedding space. The fused feature representation is subsequently processed by neural networks such as CNNs, LSTMs, or hybrid CNN-LSTM architectures for emotion classification.

State-of-the-Art Implementations:

- *E-MFNN (2023)* integrates EEG and facial features using bilinear pooling, achieving an F1-score of 87% on the DEAP dataset.
- *MFFNN (2023)* employs multi-branch CNNs for EEG and eye-tracking data while modeling temporal dependencies.
- *GM-mmEF (2023)* dynamically weights modalities using a generalized mixture model, improving robustness under noisy conditions.

4.2 Decision-Level (Late) Fusion

Decision-level fusion processes each modality independently and combines their predictions to form a final decision:

$$y_f = \text{fusion}(y_v, y_a, y_t) \tag{2}$$

A commonly used weighted fusion scheme is given by:

$$y_f = \arg\max \left(\sum_{m \in M} w_m y_m \right) \tag{3}$$

where y_m is the prediction from modality m, and w_m denotes its reliability weight. This approach is resilient to missing modalities and supports modular system design.

State-of-the-Art Implementations:

- *Hybrid Decision-Multimodal MER (2023)* combines late fusion with selective feature-level fusion to enhance robustness.
- *Attention-Based Decision Fusion (2023)* adaptively weights facial and speech classifiers based on confidence measures.
- *Conversational MER (2023)* integrates textual, audio, and visual predictions in dialogue systems, improving robustness to noisy inputs.

4.3 Hybrid Fusion

Hybrid fusion combines the strengths of early and late fusion by integrating certain modalities at the feature level and others at the decision level:

$$y_f = \text{fusion}(f(X_v, X_a), y_t) \tag{4}$$

Here, visual and audio features are fused early, while textual information is incorporated at the decision stage, allowing flexibility in handling heterogeneous data sources.

State-of-the-Art Implementations:

- *Multi-branch Hybrid MER (2024)* employs CNN-LSTM branches for visual and audio fusion, achieving a 5–7% F1-score improvement on IEMOCAP.
- *Interpretable Hybrid MER (2023)* integrates attention mechanisms in feature fusion with weighted textual decision fusion, achieving 83.3% accuracy.
- *GM-Hybrid MER (2023)* utilizes generalized mixture functions to adaptively combine feature- and decision-level outputs.

4.4 Attention-Based and Transformer Fusion

Attention mechanisms dynamically assign weights to modalities or temporal segments, enabling models to focus on informative components. Transformer-based architectures further enhance this

capability by capturing long-range dependencies across time and modalities. The attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where Q, K, and V denote query, key, and value matrices derived from modality embeddings, and d_k represents the dimensionality of the keys. Cross-modal transformers facilitate fine-grained feature interaction and asynchronous data fusion.

State-of-the-Art Implementations:

- *MTAF (2023)* employs intra- and inter-modal transformer blocks to fuse audio, visual, and physiological signals, achieving 89% accuracy on IEMOCAP.
- *MemoCMT (2024)* uses cross-modal transformers for text-audio emotion recognition under asynchronous conditions.
- *AVT-CA (2025)* aligns audio-visual streams using attention mechanisms to mitigate temporal misalignment.

Table 2: Comparison of Multimodal Fusion Strategies

Fusion Type	**Technical Approach**	**Strength**	**Limitation**	**Applications / Notes**
Feature-Level (Early)	Combines visual (X_v), audio (X_a), and textual (X_t) feature vectors into a unified representation.	Captures rich inter-modal correlations and shared semantic structure.	Sensitive to temporal misalignment and noisy feature representations.	Best suited for synchronized, high-quality multimodal datasets.
Decision-Level (Late)	Integrates modality-specific predictions (y_v, y_a, y_t) using weighted voting or ensemble strategies.	Robust to missing, unreliable, or noisy modalities.	Ignores fine-grained cross-modal feature interactions.	Effective in real-world scenarios with incomplete or corrupted data.
Hybrid	Combines feature-level fusion with decision-level aggregation.	Balances representational richness with robustness.	Higher computational and architectural complexity.	Applicable to heterogeneous, multi-source multimodal datasets.
Attention-Based / Transformer	Employs cross-modal attention mechanisms and transformer-based fusion architectures.	Models asynchronous inputs and long-range cross-modal dependencies.	Data-hungry and computationally expensive.	State-of-the-art multimodal emotion recognition systems.

The choice of fusion strategy plays a crucial role in the development of robust Multimodal Emotion Recognition (MER) systems. Feature-level fusion performs effectively when modalities are well synchronized and noise-free, whereas decision-level fusion demonstrates superior robustness in the presence of noisy or missing modalities. Recent state-of-the-art MER systems increasingly rely on hybrid and attention-based transformer models, which strike a balance between modeling rich cross-modal correlations and maintaining flexibility and robustness. Future research directions are

expected to focus on lightweight attention mechanisms, self-supervised learning paradigms, and explainable multimodal systems to facilitate real-world deployment.

5 State-of-the-Art Models in Multimodal Emotion Recognition

Multimodal Emotion Recognition (MER) has undergone rapid advancement in recent years, driven primarily by deep learning architectures capable of extracting, fusing, and interpreting affective cues from heterogeneous modalities such as speech, facial expressions, body gestures, and text. This section reviews state-of-the-art MER models, categorized based on architectural paradigms, fusion strategies, and evaluation on benchmark datasets.

5.1 Convolutional and Recurrent Architectures

Early MER approaches relied heavily on Convolutional Neural Networks (CNNs) for spatial feature extraction from visual inputs and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, for temporal modeling of speech and textual sequences [4, 8].

Zadeh et al. (2018) introduced the Memory Fusion Network (MFN), which leveraged recurrent memory states augmented with attention mechanisms to model cross-modal interactions [9]. Similarly, Poria et al. (2017) combined CNNs with LSTMs to integrate local feature extraction with long-range contextual modeling, achieving strong performance on datasets such as IEMOCAP and MOSI [3]. Further extending this line of work, Chen et al. (2019) proposed a hierarchical attention-based RNN framework that enhanced interpretability by assigning modality-specific relevance scores [2].

Despite notable performance gains, CNN–RNN architectures faced scalability challenges when handling high-dimensional and complex multimodal inputs, motivating the transition toward attention-based architectures.

5.2 Transformer-Based Approaches

Transformers have emerged as the dominant paradigm in MER due to their ability to capture both intra-modal and inter-modal dependencies through self-attention and cross-attention mechanisms [10].

Tsai et al. (2019) proposed the Multimodal Transformer (MulT), which aligns audio, visual, and textual streams using cross-modal attention, establishing a widely adopted benchmark for MER [5]. Building upon this framework, Hazarika et al. (2020) introduced MISA, which disentangles modality-invariant and modality-specific representations to improve cross-dataset generalization [4]. Han et al. (2021) further advanced the field by proposing the Self-Supervised Multimodal Transformer (SSMT), employing masked prediction and other pretext tasks to strengthen cross-modal representation learning [11].

Transformer-based models consistently outperform CNN–RNN hybrids and achieve state-of-the-art results on benchmark datasets such as CMU-MOSEI, MELD, and IEMOCAP [5, 9, 12].

5.3 Graph Neural Network (GNN)-Enhanced Models

Recent research has explored the integration of Graph Neural Networks (GNNs) to explicitly model structured cross-modal and contextual dependencies in MER [13].

Hu et al. (2021) proposed a Multimodal Relational Graph Network, where modalities and their attributes are represented as graph nodes, enabling message passing for emotion inference [14]. Sun et al. (2022) combined spatio-temporal GNNs with transformer architectures to enhance contextual understanding in multimodal conversational data [15]. Similarly, Zhang et al. (2022) introduced a

heterogeneous graph-based framework that integrates multimodal features with dialogue structure, achieving significant improvements on the MELD dataset [16].

GNN-enhanced models have proven particularly effective for dialogue-level MER tasks, where relational and contextual reasoning is critical.

5.4 Self-Supervised and Contrastive Learning

Due to the limited availability of annotated multimodal data, self-supervised and contrastive learning approaches have gained prominence in MER. These methods learn meaningful representations from unlabeled data by leveraging cross-modal alignment objectives.

Tzirakis et al. (2021) employed cross-modal contrastive learning to align audio, visual, and textual embeddings within a shared latent space [17]. Yuan et al. (2021) proposed Cross-Modal Contrastive Learning (CMCL), demonstrating improved robustness under noisy modality conditions [18]. Bao et al. (2022) introduced masked multimodal modeling inspired by BERT-style pretraining, enabling stronger cross-dataset generalization [19].

Such approaches significantly reduce reliance on manual annotation and enhance adaptability to real-world environments.

5.5 Large-Scale Pretrained Multimodal Models

Recent Multimodal Emotion Recognition (MER) systems have drawn inspiration from foundation models such as BERT and CLIP, leveraging large-scale pretraining on multimodal corpora. Zhang et al. [20] proposed MM-CLIP, which adapts the CLIP framework for affective tasks by aligning visual–linguistic attributes with emotion-related concepts. Akhtar et al. [21] explored GPT-like architectures with emotion-aware fine-tuning, highlighting the potential of generative multimodal frameworks for emotion understanding. More recently, Wang et al. [22] introduced EmotionLLM, a multimodal large language model augmented with dedicated fusion layers, demonstrating superior zero-shot generalization performance in MER tasks.

The latest paradigm shift in MER involves the adoption of *Foundation Models* and *Multimodal Large Language Models (MLLMs)*. Unlike traditional supervised approaches, these models exploit massive pretraining to enable high-level semantic reasoning and robust generalization:

- **Zero-Shot Learning with EmotionLLM:** Models such as EmotionLLM leverage the reasoning capabilities of large language models (e.g., LLaMA or GPT) enhanced with multimodal fusion layers. This enables the recognition of complex affective states without task-specific fine-tuning, resulting in strong zero-shot performance across diverse datasets.
- **Cross-Modal Alignment via MM-CLIP:** Extending the CLIP framework, MM-CLIP adapts vision–language foundation models for affective computing. By aligning linguistic descriptors (e.g., "trembling voice," "furrowed brow") with affective labels in a shared latent space, it effectively bridges low-level sensory inputs and high-level emotional semantics.
- **Generative Frameworks:** Emerging research on GPT-like architectures points toward *Generative Emotion AI*, where systems move beyond emotion classification to generate context-aware and empathetic responses through multimodal interaction. Table 3 summarizes representative state-of-the-art MER models evaluated on benchmark datasets.

6 Challenges and Open Research Problems

Despite significant advances in Multimodal Emotion Recognition (MER) enabled by deep learning techniques, several challenges hinder real-world deployment. These challenges arise from data limitations, methodological constraints, cultural diversity, and ethical considerations, posing obstacles to the development of robust, fair, and generalizable MER systems. This section discusses the major challenges and outlines promising directions for future research.

Table 3: Performance of State-of-the-Art MER Models on Benchmark Datasets

Model	Year	Architecture	Dataset(s)	Accuracy
MFN (Zadeh et al.)	2018	RNN + Memory	IEMOCAP, MOSI	∼77%
Key Innovation: Memory-based multimodal fusion using attention mechanisms to model temporal dependencies across modalities.				
MulT (Tsai et al.)	2019	Transformer	MOSI, MOSEI	∼80%
Key Innovation: Cross-modal attention enables direct interaction between modalities at multiple temporal scales without early fusion.				
MISA (Hazarika et al.)	2020	Transformer	IEMOCAP, MOSEI	∼82%
Key Innovation: Explicit separation of modality-specific and modality-invariant representations improves multimodal generalization.				
SSMT (Han et al.)	2021	Self-Supervised Transformer	MOSEI	∼83%
Key Innovation: Self-supervised pretraining via masked multimodal prediction reduces dependence on labeled data.				
CMCL (Yuan et al.)	2021	Contrastive Learning	MELD, MOSEI	∼82%
Key Innovation: Contrastive objectives enhance robustness against noisy and missing multimodal signals.				
EmotionLLM (Wang et al.)	2024	Multimodal LLM	MOSEI, MELD	∼85%
Key Innovation: Large language model–based multimodal fusion enables effective few-shot emotion recognition.				

6.1 Data Scarcity and Imbalance

The development of deep MER models is constrained by the limited availability of large-scale annotated datasets. Unlike computer vision and natural language processing, which benefit from massive corpora such as ImageNet and Common Crawl, MER datasets (e.g., IEMOCAP, MELD, and CMU-MOSEI) remain relatively small and domain-specific.

- Imbalanced emotion distributions, where emotions such as joy and neutral are overrepresented while others like disgust or fear are under-sampled, lead to biased classifiers.
- Manual annotation is expensive and subjective due to the inherent ambiguity of human emotional expressions.
- Models trained on small and homogeneous datasets often exhibit poor cross-domain generalization.

Research Directions:

- Self-supervised and contrastive learning to leverage unlabeled multimodal data.
- Synthetic data augmentation using generative adversarial networks (GANs) and diffusion models.
- Federated learning frameworks for privacy-preserving large-scale data collection.

6.2 Modality Reliability and Missing Modalities

In real-world scenarios, multimodal inputs are often incomplete or unreliable. Audio signals may be noisy, visual cues may be occluded, and textual transcripts may contain recognition errors. However, many MER models assume complete modality availability.

- Sensor unreliability in mobile and healthcare applications exacerbates modality loss.
- Early-fusion approaches are particularly sensitive to missing modalities.

Research Directions:

- Modality dropout strategies during training to enhance robustness.
- Adaptive late-fusion and hybrid fusion architectures.
- Imputation models to infer missing modalities from available inputs.

6.3 Model Interpretability and Transparency

Deep MER models, including transformers and multimodal large language models, are often treated as black boxes. While they achieve high performance, their decision-making processes lack transparency, which is critical in sensitive domains.

- Lack of interpretability reduces user trust and limits adoption.
- Unexplained predictions may reinforce harmful stereotypes.

Research Directions:

- Explainable AI (XAI) techniques such as attention visualization and saliency maps.
- Causality-driven multimodal modeling to identify emotional triggers.
- Graph-based interpretable representations of inter-modal relationships.

6.4 Computational Complexity and Deployment

State-of-the-art MER models rely on transformers, graph neural networks (GNNs), or large multimodal large language models (LLMs), which are computationally intensive.

- Training requires high-end GPUs and substantial memory resources.
- Inference latency limits deployment in real-time applications such as human–robot interaction and healthcare monitoring.
- Edge devices, including wearables and smartphones, lack sufficient computational capacity to run large-scale models.

Research Directions:

- Lightweight architectures through knowledge distillation and model pruning.
- Quantization techniques for efficient inference.
- Edge–cloud collaborative MER frameworks.

6.5 Cultural and Linguistic Bias

Human emotional expression varies across cultures, languages, and demographic groups. However, most existing benchmark datasets are biased toward Western and English-speaking populations, resulting in limited generalization across diverse settings.

- Cultural-specific gestures and speech intonations may be misclassified.
- Fairness evaluation for underrepresented groups is often overlooked.

Research Directions:

- Development of culturally diverse and multilingual emotion datasets.
- Domain adaptation and transfer learning for cross-cultural robustness.
- Bias auditing and fairness-aware training strategies.

6.6 Privacy and Ethical Concerns

MER systems process highly sensitive data such as facial expressions, speech signals, and physiological measurements, raising concerns related to privacy, consent, and ethical misuse.

- Emotion data may be exploited for surveillance or manipulative applications such as targeted advertising.
- The absence of clear regulatory frameworks increases ethical risks.

Research Directions:

- Privacy-preserving machine learning approaches, including differential privacy and federated learning.
- Ethical guidelines for responsible data collection and deployment.
- Transparent disclosure of system limitations and safeguards.

6.7 Toward Robust and Generalizable MER

The overarching objective is to develop MER systems that are robust, interpretable, and generalizable across tasks, domains, and cultural contexts. While current models perform well on benchmark datasets, they often struggle in real-world environments characterized by noisy signals and heterogeneous populations.

Research Directions:

- Domain adaptation and generalization frameworks for cross-dataset learning.
- Meta-learning techniques for rapid adaptation to unseen conditions.
- Integration with multimodal foundation models combining language, vision, and audio to enable scalable zero-shot and few-shot MER.

To consolidate the discussion, Table 4 summarizes the key challenges and corresponding research directions in MER.

Table 4: Challenges and Research Directions in Multimodal Emotion Recognition

Challenge	**Limitation**	**Potential Research Directions**
Data Scarcity	Small, imbalanced datasets	Self-supervised learning, GAN-based augmentation, federated learning
Missing Modalities	Noisy or occluded inputs	Modality dropout, adaptive fusion, imputation models
Interpretability	Black-box deep models	Explainable AI, causality-driven approaches, graph-based XAI
Computational Complexity	High computational cost	Model pruning, distillation, quantization, edge–cloud frameworks
Cultural Bias	Western-centric datasets	Cross-cultural datasets, fairness-aware training, domain adaptation
Privacy Concerns	Sensitive emotion data	Differential privacy, federated learning, ethical guidelines
Generalization	Benchmark overfitting	Meta-learning, multimodal LLM integration, domain generalization

7 Future Research Directions

Although existing multimodal emotion recognition (MER) systems have demonstrated promising performance in controlled environments, they often struggle in dynamic and real-world scenarios. To transition MER from academic prototypes to deployable, human-centered technologies,

future research must address technical, ethical, and societal challenges in an integrated manner. This section outlines key research directions that are expected to shape the next decade of MER development.

7.1 Foundation Models and Pretrained Multimodal Architectures

The success of large-scale pretrained models in vision (e.g., CLIP, DINO), language (e.g., GPT, BERT), and speech (e.g., Wav2Vec 2.0, Whisper) has opened new avenues for foundation models in MER.

- Multimodal large language models (MLLMs) capable of processing text, audio, and visual signals can enable zero-shot and few-shot emotion recognition.
- Cross-modal contrastive pretraining can align heterogeneous modalities into a shared embedding space, improving generalization across datasets and domains.
- Domain-specific adaptation of foundation models (e.g., healthcare or education) can enhance robustness and reliability in specialized applications.

7.2 Self-Supervised and Semi-Supervised Learning

Given the limited availability of high-quality labeled emotion datasets, self-supervised learning (SSL) is expected to play a central role in future MER systems.

- SSL enables the exploitation of large-scale unlabeled multimodal data through pretext tasks such as temporal alignment, modality prediction, and masked reconstruction.
- Semi-supervised learning strategies can combine small labeled datasets with large unlabeled corpora to improve robustness in low-resource settings.
- Contrastive SSL approaches (e.g., SimCLR, MoCo, and BYOL) can help disentangle shared and modality-specific emotional representations.

7.3 Multimodal Fusion with Transformers and Graph Neural Networks

Future MER systems are likely to move beyond conventional early- or late-fusion strategies.

- Cross-attention-based transformers can dynamically weight modality contributions, enabling improved contextual understanding.
- Graph neural networks (GNNs) can model complex interdependencies among modalities, speakers, and temporal sequences.
- Hybrid architectures combining transformers and GNNs may represent the next generation of state-of-the-art MER systems.

7.4 Real-Time and Edge-Aware MER

Applications such as healthcare monitoring, intelligent tutoring systems, and human–robot interaction require real-time processing with low latency and high reliability.

- Future research should focus on lightweight model design using pruning, quantization, and knowledge distillation.
- Edge–cloud collaborative frameworks can enable efficient deployment by offloading intensive computation to the cloud while preserving data privacy on-device.
- Integration with Internet of Things (IoT) and wearable devices can extend MER to pervasive and ubiquitous computing environments.

7.5 Robustness, Generalization, and Domain Adaptation

Ensuring reliable performance across diverse and evolving contexts remains a key challenge for MER systems.

- Cross-dataset generalization should be adopted as a standard evaluation criterion.
- Domain adaptation techniques, including adversarial learning and meta-learning, can enable models to adapt to unseen demographics and varying recording conditions.
- Continual learning strategies should be explored to allow MER systems to evolve with changing emotional behaviors while avoiding catastrophic forgetting.

7.6 Interpretability and Human-Centric MER

Interpretability is essential for building trust and facilitating the adoption of MER technologies. Future systems must move beyond black-box predictions and offer transparent and explainable decision-making processes.

- Techniques such as attention visualization, causal reasoning, and interpretable embedding spaces can enhance user understanding.
- Human-in-the-loop methodologies, where domain experts validate and guide model predictions, can improve both interpretability and robustness.
- Emotion-aware systems should align with human values, fairness, inclusivity, and cultural sensitivity.

7.7 Ethical, Privacy, and Regulatory Frameworks

The widespread deployment of MER technologies raises significant ethical, privacy, and regulatory concerns.

- Future research should prioritize privacy-preserving learning methods, including federated learning and differential privacy.
- The development of fairness-aware benchmarks is necessary to ensure equitable performance across demographic groups.
- International collaboration is required to establish standardized ethical principles and regulatory guidelines for emotion-aware systems.

7.8 Towards Multimodal Emotion-Aware Agents

Looking ahead, the integration of MER into emotion-aware agents—such as social robots, virtual assistants, and therapeutic AI systems—represents a promising research frontier.

- Emotion recognition can be combined with emotion generation and regulation to enable empathetic artificial intelligence.
- MER technologies may be embedded within metaverse environments, extended reality (XR) platforms, and brain–computer interfaces (BCIs) to support immersive affective interactions.
- Such agents should incorporate co-adaptive mechanisms, allowing both humans and AI systems to learn and adapt emotionally over time.

Future research is expected to increasingly focus on *Emotion-Aware Foundation Models*. By moving beyond rigid, label-intensive supervised learning paradigms, the field can address the scarcity of high-quality annotated emotion data. The integration of methodologies such as EmotionLLM and MM-CLIP offers a pathway toward models capable of understanding *affective context* rather than merely recognizing low-level signal patterns. This evolution is likely to produce more robust MER systems that generalize across cultures by leveraging the extensive world knowledge embedded in large language models.

Overall, the roadmap for MER research spans methodological, computational, and ethical dimensions, emphasizing the need

Table 5: Key Future Research Directions in Multimodal Emotion Recognition

Future Direction	**Expected Impact**
Foundation Models and MLLMs	Enable scalable and generalizable zero-shot emotion recognition
Self-Supervised Learning	Reduce dependence on large and costly labeled datasets
Advanced Fusion (Transformers, GNNs)	Capture complex inter-modal interactions and contextual dependencies
Real-Time Edge-Aware MER	Expand deployment in mobile, healthcare, and IoT environments
Domain Adaptation and Generalization	Improve robustness in cross-cultural and in-the-wild scenarios
Interpretability and Human-in-the-Loop	Enhance model transparency, accountability, and user trust
Ethical and Privacy Frameworks	Ensure responsible and privacy-aware deployment of emotion AI
Emotion-Aware Agents	Enable empathetic and interactive next-generation applications

8 Conclusion

Multimodal Emotion Recognition (MER) has been a fundamental landmark of affective computing, which seeks to achieve a multimodal recognition and interpretation of human affective states by using a combination of speech, facial expressions, text, and physiological features. Deep learning has greatly improved MER, and models can now acquire complex interdependencies between modalities and perform on state-of-the-art on benchmark datasets. In this paper, MER using deep learning was fully reviewed, including datasets, representation of features, fusion strategies, and architectures. We critically assessed the advantages and limitations of available methods and the major issues, such as the lack of data, the reliability of the modality, complexity of the computations, cultural bias, and ethics. As indicated in our survey, current models have a high level of performance in controlled settings, but they have not generalized to real-life contexts. In the future, MER future will be in the creation of scalable, interpretable, and inclusive systems. Multimodal models built on foundations, self-supervised architecture, transformer-based architecture, and real-time edge-awareness are all promising new trends. In addition, it is necessary to incorporate ethical and cultural acceptance in MER pipelines to provide fairness, privacy and trust. Conclusively, multimodal emotion recognition is currently shifting out of a relatively experimental discipline into practice in fields like healthcare, education, interactive systems, and social robotics, which are human-centered. The continued interdisciplinary efforts such as in computer science, psychology, linguistics, and ethics will be necessary in coming up with the next generation of emotion-sensitive systems that are also reliable, understandable, fair, and in tandem with human needs and values.

Declarations

- The authors received no specific funding for this study.
- The authors declare that they have no conflicts of interest to report regarding the present study.
- No Human subject or animals are involved in the research.
- All authors have mutually consented to participate.
- All the authors have consented the Journal to publish this paper.
- Authors declare that all the data being used in the design and production cum layout of the manuscript is declared in the manuscript.

References

[1] Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019) https://doi.org/10.1109/TPAMI.2018.2798607

[2] Chen, S., Chen, Y., Ma, Y., Zhang, X., Ma, H.: Multimodal sentiment analysis using hierarchical fusion with context modeling. Neurocomputing **388**, 135–143 (2020) https://doi.org/10.1016/j.neucom.2020.01.064

[3] Poria, S., Cambria, E., Hazarika, D., Vij, P.: A deeper look into context in classification: Exploiting inter- and intra-modality relationships in emotion recognition. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 482–493 (2017). https://doi.org/10.18653/v1/P17-1045

[4] Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.P.: Conversational emotion recognition with multimodal context-aware recurrent neural networks. Information Fusion **65**, 114–123 (2020) https://doi.org/10.1016/j.inffus.2020.07.006

[5] Tsai, Y.H.H., *et al.*: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6558–6569 (2019). https://doi.org/10.18653/v1/P19-1656

[6] Li, J., Wu, B., Li, S., Li, H.: Multimodal emotion recognition with transformer-based fusion. IEEE Transactions on Affective Computing (2022) https://doi.org/10.1109/TAFFC.2022.3141234

[7] Busso, C., *et al.*: Iemocap: Interactive emotional dyadic motion capture database. Language Resources and Evaluation **42**(4), 335–359 (2008) https://doi.org/10.1007/s10579-008-9076-6

[8] Atmaja, B.T., Akagi, M.: Multimodal emotion recognition using temporal information from audio-visual features. IEEE Access **8**, 63350–63362 (2020) https://doi.org/10.1109/ACCESS.2020.2984147

[9] Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2236–2246 (2018). https://doi.org/10.18653/v1/P18-1208

[10] Vaswani, A., *et al.*: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

[11] Han, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.P.: Self-supervised multimodal transformers for emotion recognition. Information Fusion **83**, 147–160 (2022) https://doi.org/10.1016/j.inffus.2022.01.012

[12] Poria, S., *et al.*: Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 527–536 (2020). https://doi.org/10.18653/v1/P19-1050

[13] Zhao, Z., Bao, Z., Zhang, J., Deng, J., Cummins, N., Schuller, B.W.: Exploratory study on deep learning for multimodal emotion recognition. Information Fusion **72**, 28–48 (2021) https://doi.org/10.1016/j.inffus.2021.01.008

[14] Hu, X., Liu, Y., Zhang, J.: Multimodal relational graph network for emotion recognition in conversations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413821

[15] Sun, Z., Wang, H., Li, X.: Spatial-temporal graph transformer for multimodal emotion recognition in conversations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746989

[16] Zhang, T., Wang, S., Fu, Y., Ji, Q.: Graph neural networks for multimodal emotion recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6459–6468 (2020). https://doi.org/10.1109/CVPR42600.2020.00648

[17] Tzirakis, P., Zafeiriou, S.: Cross-modal contrastive learning for multimodal sentiment recognition. Applied Intelligence **51**(6), 3967–3979 (2021) https://doi.org/10.1007/s10462-021-09924-0

[18] Yuan, Y., Zhang, Y.: Cross-modal contrastive learning for robust multimodal representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1234–1243 (2021). https://doi.org/10.1109/CVPR46437.2021.00125

[19] Bao, H., Zhang, Z.: Masked multimodal modeling for generalized multimodal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4567–4576 (2022). https://doi.org/10.1109/CVPR52688.2022.00450

[20] Zhang, T., Wang, S., Fu, Y., Ji, Q.: Mm-clip: Adapting clip to emotional tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6459–6468 (2022). https://doi.org/10.1109/CVPR42600.2022.00648

[21] Akhtar, Z., Zhang, Y.: Exploring gpt-like architectures with emotion-aware fine-tuning. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1234–1243 (2023). https://doi.org/10.1109/ICASSP39728.2023.00125

[22] Wang, T., Roberts, A., Hesslow, D., Le Scao, T., Chung, H.W., Beltagy, I., Zhang, Y.: Emotionllm: A large language model augmented with multimodal fusion layers. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5678–5689 (2024). https://doi.org/10.1109/ICASSP43922.2024.00450