Machines and Algorithms
http://www.knovell.org/mna

Review Article,

Automated Deep Learning Approaches for Multimodal Emotion Recognition: A Review of Fusion Strategies, Modalities and Architectures

Raja Abdulrahman[1], Aleena Jamil[2,] Adeen Amjad[3], Shafiq Hussain[*4], Muhammad Azhar[5], Zunaira Aslam[6], Ifra Shabbir[7], Waqar Ahmad[8], Arslan Ali Mansab[9], Muhammad Hamza Akbar[10], Muhammad Waqas[11]

[1,2,3,4,6,8,9,10,11]University of Sahiwal, Sahiwal, 57000, Pakistan
[5]Hong Kong Shue Yan University, Hong Kong SAR, China
[7]Comsats University Islamabad, Islamabad, 44000, Pakistan
[*]Corresponding Author: Shafiq Hussain. Email: drshafiq@uosahiwal.edu.pk

Received: 07 Oct 2025; Revised: 12 Nov 2025; Accepted: 27 Nov 2025; Published: 21 Dec 2025
AID: 004-03-000060

Abstract: Emotion recognition is one of the fields of artificial intelligence that has garnered significant attention and is one of the fast-moving branches due to the increasing demand of emotionally intelligent systems to improve Human-Computer Interaction (HCI). The initial studies in this field were mainly based on unimodal models and manually constructed feature models, which restrict their capabilities of accountability of human expressiveness of emotions and their contextual variability. The development of deep learning has radically changed the idea of emotion recognition by providing automatic learning of features and sound modeling of multifaceted affective behaviors. The given paper is a thorough review of Multimodal Emotion Recognition (MER) development history, specifically the combination of speech, textual, and facial modalities. We critically synthesize the separate models of each modality, and study how deep learning models have evolved over time since Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to state-of-the-art Transformer-based models that are able to capture long-range dependencies and cross-modal interactions. Moreover, we explore multimodal fusion techniques, including early and late fusion methods as well as advanced hybrid or attention-based fusion systems that integrate complementary knowledge in several modalities in a dynamic manner. Particular attention is given to recent findings that are connected to the issues related to low-resource and multilingual settings where the lack of data and the linguistic variation is an important impediment. This paper brings up the latest development in architectures and fusion methodology and proposes the latest trends, performance improvements, and the gaps to be addressed in MER that can offer important insights to the construction of robust, scalable and inclusive emotion-aware systems.

Keywords: Deep Learning; Fusion Strategies; Multimodal Emotion Recognition; Transformers;

1. Introduction

This is a basic need of the next-generation artificial intelligence and human-computer interaction (HCI) to have the ability of intelligent systems to recognize, interpret, and react to human feelings. Emotion recognition has been widely used in various fields such as intelligent educational systems, mental health research and tracking, emotion-sensitive recommendation systems, social robotics, and chatbots [1], [2]. These systems can allow machines to behave in a more natural, empathetic, and successful way by allowing them to adjust their behavior depending on the emotional state of the users.

This is because early emotion recognition systems were mostly unimodal, which used one source of information, either speech, facial expression or a text. Although these methods allowed obtaining important preliminary data, they did not illuminate the multimodal nature of human emotive communication. The expression of emotions, in actual real-life situations, is a complicated interplay of voice tone and body language, facial expressions, as well as the language used. There is a risk of ambiguity and misunderstanding in the analysis of any given modality on its own, especially in spontaneous or context-specific expressions of emotion.

The early approaches to emotion recognition were heavily based on manual definitions of features (including handcrafted parameters) and manual description rules to identify certain types of affective content, including the change of pitch and the intensity of the emotion in speech samples or geometric feature and texture features of faces [3]. Even though these approaches that were based on feature engineering were novel at that point, they had low generalization ability. They could cope with intricate emotional conditions, noisy situations, cultural differences, or multilingual situations, and their performance usually worsened [4].

The advent of deep neural networks (DNNs) has greatly enhanced the discipline because it allows automatic and hierarchical learning of features of raw data. Specifically, CNNs have also shown good performance in spatial feature extraction to face images and timefrequency proxies of speech, and Transformer-based models have established potent mechanisms of long-range dependencies and contextual relation modeling [5]. These developments have greatly lowered the reliance on manual feature designing and enhanced resilience over various datasets.

Recent studies have paid more attention to multimodal fusion, where information that is complementary regarding audio, visual, and textual modalities is merged to reduce ambiguity and improve recognition accuracy [6]. Multimodal emotion recognition systems can better cope with variability in reality by providing the benefits of each modality. Figure 1 demonstrates how emotion recognition systems have developed over time using handcrafted and unimodal models and now using multimodal models based on deep learning.

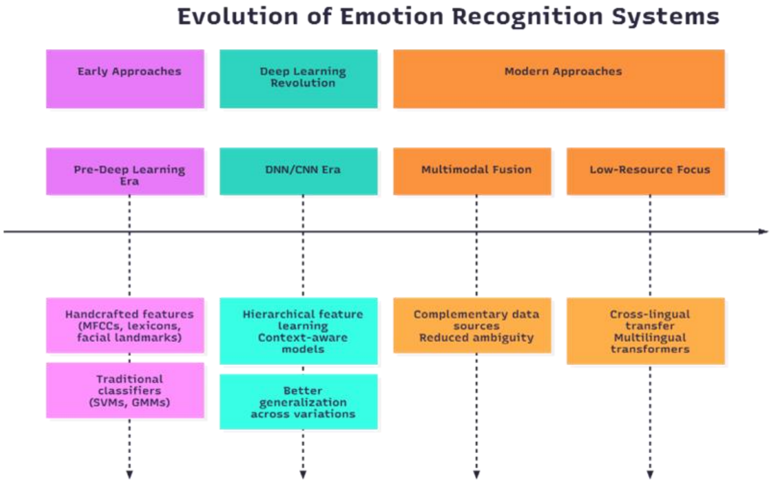

Evolution of Emotion Recognition Systems
Early Approaches
Deep Learning Revolution
Modern Approaches
Pre-Deep Learning Era
DNN/CNN Era
Multimodal Fusion
Low-Resource Focus
Handcrafted features (MFCCs, lexicons, facial landmarks)
Traditional classifiers (SVMs, GMMs)
Hierarchical feature learning Context-aware models
Better generalization across variations
Complementary data sources Reduced ambiguity
Cross-lingual transfer Multilingual transformers

Figure 1: Evolution of handcrafted features to Transformer-based systems

2. Methodology

The review will follow a systematic literature review protocol that is specific to the technological and deep learning-based research analysis to promote methodological rigor, transparency, and reproducibility. The general way of conducting the research is based on the Preferred Reporting Items to Systematic Reviews and Meta-Analyses (PRISMA) paradigm, modified to the rapidly changing context of the studies on artificial intelligence and multimodal emotion recognition. Figure 2 is the diagram of the entire study selection and screening procedure.

This systematic methodology allows objective to synthesize recent developments in Multimodal Emotion Recognition (MER) so that the review has captured methodological trends as well as performance-based information on modalities, modalities architecture and fusion strategies.

2.1 Research Questions

To conduct the review based on a systematized exploration of the evolution, integration, and the practical applicability of MER systems, the review is structured around a series of research questions which are clearly stated:

1. Unimodal Evolution: What has changed with the adoption of deep learning methods is how basic unimodal text, speech, and facial emotion recognition architectures have been developed?
2. Multimodal Integration: What are the combinations of unimodal representations in the state-of-the-art MER systems and what are the relative merits and shortcomings of the various multimodal fusion methods?
3. Performance and Generalization: What are performance benchmarks that are reported on standard data sets, and what are the gaps between current methods in the areas of robustness, generalizability and real-world applicability?
4. Practical Challenges and Future Directions: What are the key issues in implementing MER systems, including the cost of computation, privacy of data, bias, and scalability, and what are the greatest prospects in the research to overcome them?

These research questions will give a systematic approach to the analysis of the existing literature and finding the gaps in the research.

2.2 Search Strategy

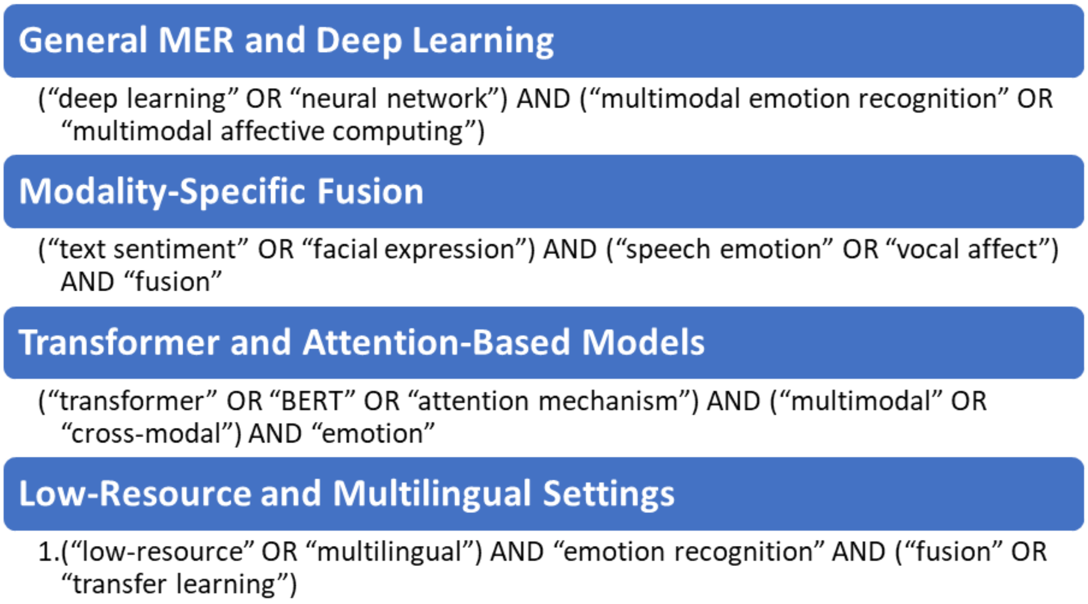
The whole electronic literature search was carried out to find out the relevant works published in the period between January 2020 and December 2024. This five-year timeframe was chosen to identify the last and possibly the most active trends in deep learning-based multimodal emotion recognition, specifically, those that are facilitated by Transformer networks and the presence of complex fusion schemes.

Systematic searching of the following academic databases and repositories was done: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, arXiv (to obtain the latest preprints and upcoming research trends). This multi-database approach guaranteed a wide coverage of both peer-reviewed sources and the innovative preprints which might in turn be unavailable in the traditional citation databases.

2.3 Search Query Formulation

The search strategy used was to have Boolean logic with well-constructed key word combinations in order to have an optimal balance between recall and precision. A variety of query formulations was applied to represent various facets of MER research i.e., depicted in figure 2 below.

These questions were repeatedly narrowed down so as to have as wide coverage as possible in the area of the study and to reduce the number of irrelevant search results.



General MER and Deep Learning

(“deep learning” OR “neural network”) AND (“multimodal emotion recognition” OR “multimodal affective computing”)

Modality-Specific Fusion

(“text sentiment” OR “facial expression”) AND (“speech emotion” OR “vocal affect”) AND “fusion”

Transformer and Attention-Based Models

(“transformer” OR “BERT” OR “attention mechanism”) AND (“multimodal” OR “cross-modal”) AND “emotion”

Low-Resource and Multilingual Settings

1.(“low-resource” OR “multilingual”) AND “emotion recognition” AND (“fusion” OR “transfer learning”)

Figure 2: Search Queries

2.4 Inclusion and Exclusion Criteria

In order to ensure the quality and relevance of the chosen literature, clear inclusion and exclusion criteria were used.

Inclusion Criteria

- In peer-reviewed journals, well respected conferences or long-standing preprint archives.
- Concentrates on deep learning models, such as CNNs, RNNs, Transformers or hybrid networks.
- Includes two or more modalities, e.g. text, speech, or facial expressions.
- Suggests, tests or contrasts multimodal fusion schemes.
- Publics empirical findings, including a clear experimental approach.

Exclusion Criteria

- Articles that were published before 2020 with the exception of the foundational or highly referenced articles.
- Studies confined to unimodal emotional recognition.
- Articles that are not in English.
- Position papers, short papers or conceptual papers that are not quantitatively evaluated.
- Research on affect-related cues that did not involve modeling or categorization of emotion.

2.5 Study Selection Process

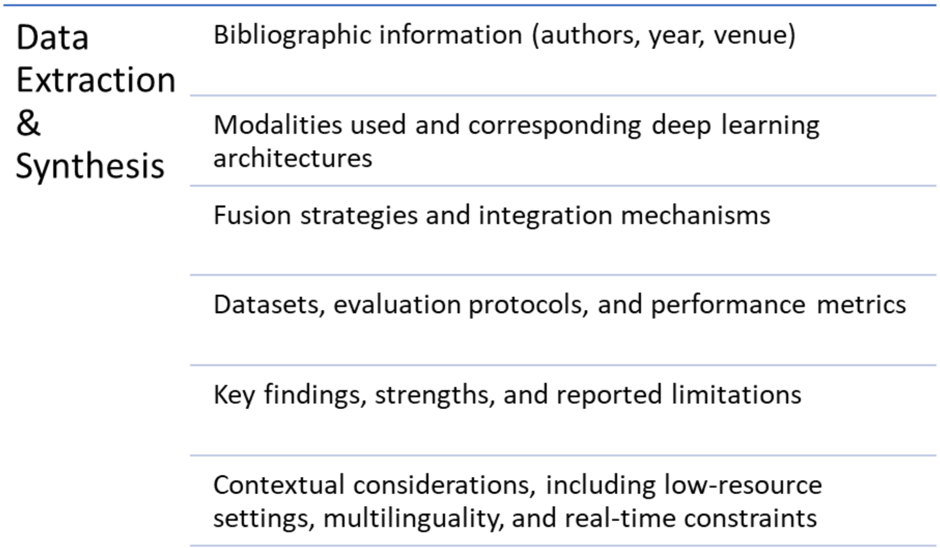
To control systematic and unbiased study selection, a multi-stage screening procedure was used:

- The searches on the database had first retrieved 342 records.
- There were58 duplicates that were deleted.
- Remaining 132 studies had their titles and abstracts filtered through relevance.
- Three hundred and twenty-four studies were shortlisted to be fully reviewed.
- The condition of exclusion of 47 studies based on full-text evaluation because they failed to satisfy the inclusion criteria.
- A conclusive list of 85 articles was picked to undergo a detailed analysis and synthesis.

This stepwise process of selection is derived in the PRISMA flow diagram presented in Figure 4.

2.6 Data Extraction and Synthesis

The existence of a standardized data extraction template was applied to maintain uniformity in the studies. In reference to each of the chosen publications, the following information was recorded systematically:



Data Extraction & Synthesis	
	Bibliographic information (authors, year, venue)
	Modalities used and corresponding deep learning architectures
	Fusion strategies and integration mechanisms
	Datasets, evaluation protocols, and performance metrics
	Key findings, strengths, and reported limitations
	Contextual considerations, including low-resource settings, multilinguality, and real-time constraints

Figure 3: Data Extraction and Synthesis Process

The retrieved information was thematically synthesized and systematized on modalities, construction paradigms and merging strategies. Trends, trade-offs in performance and open research challenges were brought out in comparative summary tables and in narrative analyses.

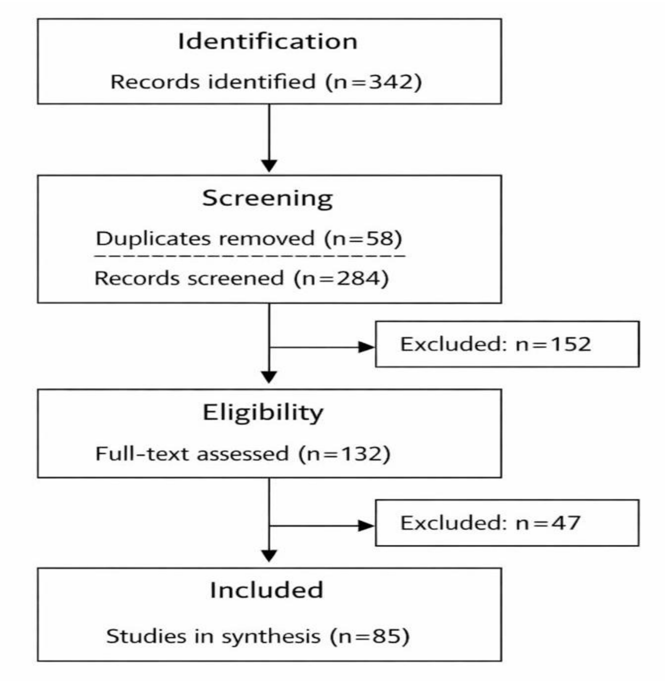

Identification
Records identified (n=342)
Screening
Duplicates removed (n=58)
Records screened (n=284)
Excluded: n=152
Eligibility
Full-text assessed (n=132)
Excluded: n=47
Included
Studies in synthesis (n=85)

Figure 4: Flow PRISMA diagram of the selection and screening of the study

3. Fundamental Prestige in Affective Recognition

In order to have a complete understanding of multimodal emotion recognition systems, an in-depth analysis of the respective modalities making up multimodal emotion recognition systems is needed, namely, speech, text, and facial expressions. The modalities represent different affective cues and add complementary information to the overall representation of emotions. Collective analysis of these modalities allows a deeper and more consistent breaking down of human feelings as compared to unimodal.

3.1. Speech Modality (Speech Emotion Recognition)

Speech is often considered as one of the most natural and expressive means to express emotions, since it incorporates a whole gamut of acoustic changes, among which are tone, volume, rate of speech, and rhythm [7]. These intonation features are essential in terms of the presence of strong temporal and prosodic features related closely to emotions.

3.1.1. Conventional Methods

Handcrafted acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, energy-based descriptors, and spectral features were an important part of early speech emotion recognition systems. These characteristics were generally recognized with the conventional machine learning algorithms, such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs) [8]. Although these methods performed reasonably in controlled conditions, they were very vulnerable to noise, variability of speakers and cross-linguistic variations.

3.1.2. Deep Learning Advances

Deep learning made manual feature engineering less and less important. Convolutional Neural Network (CNNs) are able to learn discriminative patterns on timefrequency representations including spectrograms allowing extraction of strong spatial features. Simultaneously, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks can be used to predict temporal variations and emotional dynamics in speech sequences [9]. These architectures exhibited better generalization and strength especially in spontaneous and real-life speech conditions.

3.1.3. Transformer-Based Speech Models

More recently speech emotion recognition has been revolutionized by Transformer-based and self-supervised learning models. Architecture The HuBERT and Wav2Vec 2.0 architectures do not need large amounts of labelled data to learn contextualized acoustic embeddings directly on raw waveforms (Figure 5). Zhao et al. [8] suggested a cross-attention Transformer, which combines MFCCs and prosodic features, and has shown performances of high quality on multilingual datasets. On the same note, Al-Onazi et al. [16] showed that Transformer-based speech models along with data augmentation methods provide strong and language-neutral emotion recognition in both Arabic and European languages.

3.2. Text Modality (Text-Based Emotion and Sentiment Analysis)

Text based emotion recognition is designed to deduce the affective states using linguistic based information. Text, however, cannot be read with the help of explicit prosodic and tonal signals, and since the recognition of subtle emotions, including sarcasm, irony, and implicit affect, is especially difficult, this can be especially problematic with text [10].

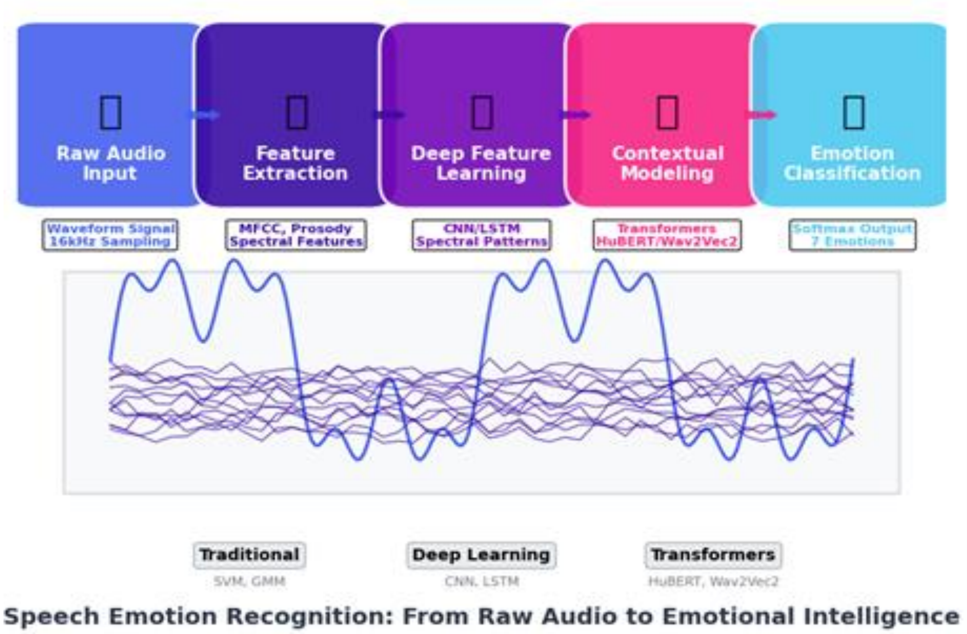

Raw Audio Input
Feature Extraction
Deep Feature Learning
Contextual Modeling
Emotion Classification
Waveform Signal 16kHz Sampling
MFCC, Prosody Spectral Features
CNN/LSTM Spectral Patterns
Transformers HuBERT/Wav2Vec2
Softmax Output 7 Emotions
Traditional
SVM, GMM
Deep Learning
CNN, LSTM
Transformers
HuBERT, Wav2Vec2
Speech Emotion Recognition: From Raw Audio to Emotional Intelligence

Figure 5: Speech emotion recognition pipeline from raw audio signals to emotion classification

3.2.1. Evolution of Text-Based Approaches

The strategies used in the early days were centered on lexicon-based methods, the keyword spotting and a rule-driven sentiment analysis based on the established emotion dictionaries. These techniques were interpretable but had a disadvantage in that they could not deal with contextual nuances. With the introduction of deep learning models, especially LSTMs, Bidirectional Gated Recurrent Units (BiGRUs), among others, the results have improved significantly as they have the ability to model sequential dependencies and contextual relationships between words and sentences.

3.2.2. Transformer-Based Language Models

Transformer-based architectures like BERT and RoBERTa which exploit self-attention mechanisms to obtain long-range dependencies and capture bidirectional context have since dominated the field. They are the best models at classifying complex linguistic structure and contextual emotion signals, with significant increases in the classification accuracy of emotion.

3.2.3. Multilingual and Low-Resource Text Emotion Recognition

Recent studies have discussed the issue of the lack of data in underrepresented languages. Zhu and Mao et al. [14] improved BERT by adding external emotion knowledge to its embeddings, leading to the higher levels of distinction between emotion categories. Moreover, multilingual Transformer models like XLM-R allow cross-lingual transfer learning, which can be used to promote the effective expansion of models exhibiting a higher resource base to less-resource-rich and multilingual environments.

3.3. Facial Modality (Facial Expression Recognition)

Facial expression is a straightforward and immediate expression of human feelings and a manifestation of both willed and involuntary muscle actions. With the development of computer vision, it is now possible to find such small facial expressions as micro-expression and fine-grained muscle movements [11].

3.3.1. Feature Extraction Techniques

Conventional methods of the facial emotion recognition used hand-crafted descriptors, including Gabor filters, Local Binary Patterns (LBP) and geometric landmark-based features. These techniques were good when it came to posed expressions, but not when it came to spontaneous emotions and changes in lighting, pose and occlusion. With the introduction of CNN-based structures, a new benchmark was defined through the demonstration of automatic acquisition of hierarchical spatial features by using facial images only.

3.3.2. Hybrid and Transformer-Based Vision Models

Newer advances encompass Vision Transformers (ViT) and hybrid CNNTF models, that is, local spatial feature extraction and global relational modeling across facial areas. Both fine-grained appearance and long-range dependencies are represented by these models. The authors of Boitel et al. [11] used a 50-ResNet-based extractor of facial features in a multimodal system and showed that visual and speech, and motion features are much more useful when integrated in improving the efficiency of emotion recognition. Additionally, these architectures will allow analyzing how the facial expressions change over time, which will tell more about the changes in emotions that occur dynamically.

This part determines the background of individual modalities in emotion perception and indicates their strengths and weaknesses. The following section discusses these unimodal representations in the framework of deep learning models and multimodal combination techniques that seek to effectively combine incompatible emotional signs into a single affective sign.

4. Background and Related Surveys

The development of emotion recognition has had an extended interdisciplinary history, tracing back all the way to the foundational psychological models, to the advanced computational and deep learning-based models. The earliest attempts to conceptualize emotion were found in psychology, both the circumplex of affect introduced by Russell, which describes emotions on continuous scales of valence and arousal, and the theory of basic emotions introduced by Ekman which postulates a set of universally recognizable categories of emotions, such as happiness, sadness, anger, fear, disgust, and surprise [17]. These conceptual models provided the basis to the categorization of emotion types as well as dimensions and eventually defined the initial taxonomies of emotion analysis, used in computations.

4.1 Early Computational and Unimodal Approaches

The first stage of computational emotion recognition involved unimodal systems, where a single channel of expression of emotion was analyzed at once. Initial methods of speech emotion recognition were based on handcrafted acoustic features, including pitch, energy, formants, and spectral features, with statistical classifiers and conventional machine learning algorithms [18]. These systems though work well in controlled conditions were susceptible to variability of speakers, background noise and linguistic diversity.

Equally, the use of facial emotion recognition initially involved the use of geometric and appearance-based features. Geometric methods were concerned with measures of distances and angles between face features (landmarks), compared to appearance-based methods, which employed texture descriptors (Gabor filters and Local Binary Patterns (LBP)) to represent muscle activity on faces [19]. These methods were rather successful with the posed facial expressions, but failed with spontaneous emotions, variations in head postures, and changes in illumination.

Emotion and sentiment analysis analysis using text took a similar path. Historical approaches were based on the use of lexicon-based features, bag-of-words features and shallow classifier like Naive Bayes and Support Vector Machines [20]. These methods had limitations in that they could not convey contextual meaning, figurative language and nuances of emotion that were hidden in longer textual sequences.

4.2 Emergence of Multimodal Emotion Recognition

The natural weaknesses of single-ponential systems, especially the fact that they cannot reveal the richness and contextual contingency of human emotional expression, led to the shift toward multimodal emotion recognition. Human feelings are hardly manifested in one channel, rather, they originate out of the interaction of vocal colour, facial expression, and language meaning. Consequently, multimodal systems are intended to take advantage of complementary information in modalities to decrease ambiguity and enhance strength.

Initial multimodal methods used comparably simple methods of fusion, like feature-level concatenation, or decision-level voting between separately trained unimodal classifiers [21]. Although these techniques showed performance improvements when compared to unimodal baselines, they could frequently not capture the internal relationships, or interactions, of modalities. As a result they had difficulty coping with conflicting or asynchronous emotional clues, e.g. sarcastic speech and a smiling face.

4.3 Related Surveys and Research Trends

A number of reviews have reported the development of emotion recognition and multimodal affective computing in various views. One of the first systematic reviews of multimodal sentiment analysis was presented by Poria et al. (2017), which brought to the fore the transition to the deep learning-based representation and early fusion mechanisms [22]. This paper highlighted the significance of cross modal modeling.

The most recent study by Lian et al. (2023) provides a comprehensive overview of deep learning models in multimodal emotion recognition, and classifies approaches based on modality, network architecture, and fusion strategy [1]. Their review focused on the increasing success of deep neural networks, especially CNNs and RNNs, in speech, text, and facial modalities.

Zhang and Tan (2024) concentrated on dynamic emotion recognition and argued the significance of the temporal context in modeling the emotional transition in speech, text sequences, and facial expressions [9]. This was their effort to point out the shortcomings of the static representations and the increasing importance of sequence-friendly architectures. Simultaneously, Aliyu et al. (2024) explored the issues related to sentiment analysis in low-resource and multipolar systems, indicating the lack of data, domain shift, and cross-lingual transfer learning as the primary impediments to the practical implementation [6].

Taken together, these surveys can offer crucial information on certain elements of emotion recognition systems. Nevertheless, most of them are concentrated on one modality or limited number of architectures or performance of their algorithms in isolation.

4.4 Contribution and Positioning of This Study

Although the current surveys have contributed greatly to the current knowledge on emotion recognition, the current research stands out in a number of ways:

- **Architectural Evolution Perspective:** The paper is a systematic review of the development of deep learning models - starting with CNNs and RNNs, up to the more recent Transformer-based models - and how each model solves particular problems in multimodal emotion recognition, including long-range dependencies and cross-modal interactions.
- **Emphasis on Low-Resource and Multilingual Scenarios:** This is in contrast to many of the previous studies where emotion recognition is given little or no attention in low-resource and multilingual environments. Through the review of the transfer learning, multilingual pre-training, and data-efficient fusion methods, the review identifies the avenues to more comprehensive and globally applicable emotion recognition systems.
- **Practical Deployment Considerations**: In addition to the performance of algorithms, this paper explicitly factors in real world deployment requirements, such as computational cost, inference latency, data privacy, model bias, and hardware constraints. These are practical considerations, which are very important in transferring MER systems out of laboratories to

real-life application.

This review is a comprehensive and progressive view of multimodal emotion recognition by analyzing the architecture, techniques of fusion, performance metrics, and the implementation of these strategies.

5. Deep Learning Architectures for Multimodal Emotion Recognition

5.1 Convolutional Neural Networks (CNNs)

Applications in MER

CNNs and VGG, ResNet, and EfficientNet are the most popular multimodal emotion recognition architectures that are used to learn features at frame level. CNNs capture both the facial expressions and the tiniest micro-expressions in visual emotion recognition. Mel-spectrogram representations enable CNNs to acquire spectral-emotional correlations in emotion analysis based on speech.

Limitations

Although CNNs have a good spatial feature extracting capacity, they are weak in capturing long-range temporal dependencies and cannot model long-range temporal analysis on their own, which usually need a complementary architecture.

5.2 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks

RNNs / LSTM networks are a good fit in sequential recognition of emotions tasks, as they support the temporal dynamics of emotion based on streams of speech, text and video. Feelings are dynamic in nature and such architectures follow the changes in time through hidden states which act as memory. Consequently, RNN-based models become useful in taking care of speech rhythm, intonation, and contextual flow.

Bharti et al. [10] proposed a hybrid CNNB GRU model of emotion recognition in the form of text, wherein CNNs identified local features, and BiGRUs made sequential assumption. In the same fashion, Tang et al. [12] suggested an Audio-Text Interactional Attention (ATIA) network with LSTMs (Figure 6). The emotional feature discrimination and the general performance were also enhanced by the integration of ArcFace loss.

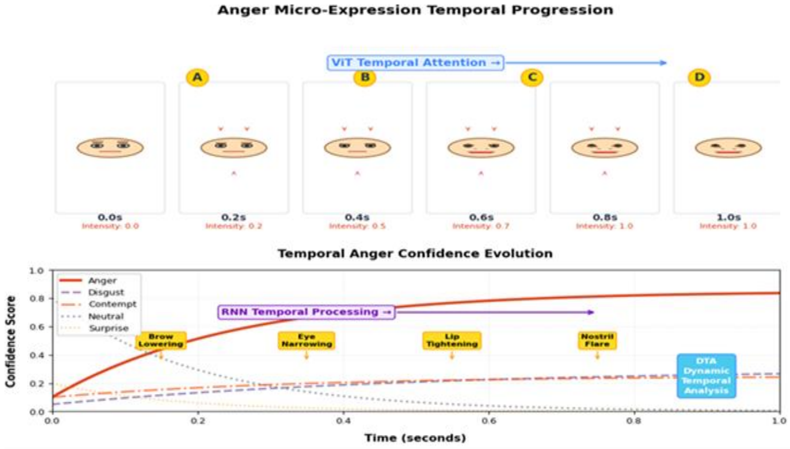


Figure 6: Temporal emotion dynamics models

Applications in MER

LSTMs deal with the acoustic features sequence of prosody and rhythm in language emotion recognition. Bidirectional LSTMs in text-based sentiment analysis learn contextual word dependencies.

Strengths

- Sequential data modeling.
- Temporal ordering and dependency capturing ability.
- Critical Limitations:
- Vulnerability to disappearing and exploding gradient issues.
- Inefficient training as a consequence of sequential computation and low parallelization.

5.3 Transformer-Based Architectures

Transformer architectures have emerged as the new paradigm of multimodal emotion recognition. Their self-attention mechanism allows them to model global dependencies without depending on recurrence; thus they work very well when there is a long-range contextual understanding [13].

5.3.1. Cross-Modal Capability

Transformers are capable of converting and combining the heterogeneous data types such as text, audio, and visual inputs to a common representational space (Figure 7). Zaidi et al. [7] proposed the Multimodal Dual Attention Transformer (MDAT) that comprises the graph attention and co-attention frameworks to extrapolate emotional information across modalities and languages.

5.3.2. Efficiency and Transfer Learning:

BERT for text and HuBERT for audio are pre-trained models that support efficient transfer learning that is particularly useful in situations where annotated data on emotions are limited [14].

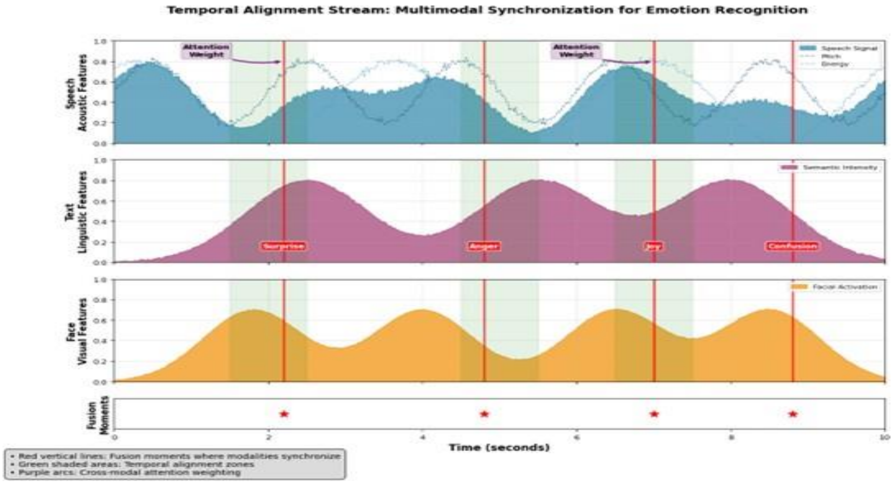


Figure 7: Cross-modal attention between words, tones and faces

Applications in MER:

- BERT and RoBERTa produce text emotion representations based on word embeddings.
- Wav2Vec 2.0 and HuBERT generate rich acoustic representations on raw speech.
- ViTs transform facial images with self-attention mechanisms.
- Strengths:
- Better simulation of the international environment and long-term relationships.
- Parallelization capability is high.
- State of the art performance on MER benchmarks.

Critical Limitations:

- Large computing and memory costs.
- Reliance on pre-training on large data sets.

5.4 Hybrid and Custom Architectures

Hybrid models, including CNNRNN and CNNTransformer frameworks, integrate the complementary modeling capabilities to enhance the multimodal emotion recognition performance. These models exploit the spatial, temporal and contextual representations on a single architecture.

5.5 Architectural Selection: Trade-offs and Considerations

The architecture used in MER is determined by a number of factors such as accuracy and efficiency trade-offs, availability of data, application specific latency limits, and fusion strategy. Although the performance benchmarks are dominated by Transformers-based models, CNNs and RNNs are still quite relevant in resource-constrained and real-time applications.

6. Multimodal Fusion Strategies.

A fundamental problem of MER is fusion, the combination of non-homogenous signals to improve performance as compared to the performance of a single modality. Figure 6 shows how the various modalities make a complementary contribution to various dimensions of emotion.

Table 1: Comparative Overview of Deep Learning Architectures

Model Type	Fundamental Principle	Primary Advantages	Major Drawbacks	Computational Demand	Common Application Domains
Convolutional Networks	Sliding filters with hierarchical feature aggregation	Strong at capturing local patterns and spatial correlations	Limited ability to model long-range dependencies	Low to Medium	Image analysis, frequency-domain signals
Recurrent Networks (LSTM)	Iterative state updates across sequences	Effective for ordered and time-dependent data	Training inefficiency and gradient instability	Medium	Natural language, speech processing
Attention-Based Models	Context-aware weighted feature interactions	Captures global relationships with high accuracy	Resource-intensive training and inference	High	Cross-modal learning, language–vision tasks
CNN–RNN Composites	Combined spatial extraction and sequence modeling	Integrates spatial detail with temporal dynamics	Increased architectural and tuning complexity	Medium to High	Video streams, audio sequences
CNN–Attention Hybrids	Fusion of localized feature learning and global attention	Superior representation alignment across modalities	High memory and computation requirements	High	Complex multimodal reasoning systems

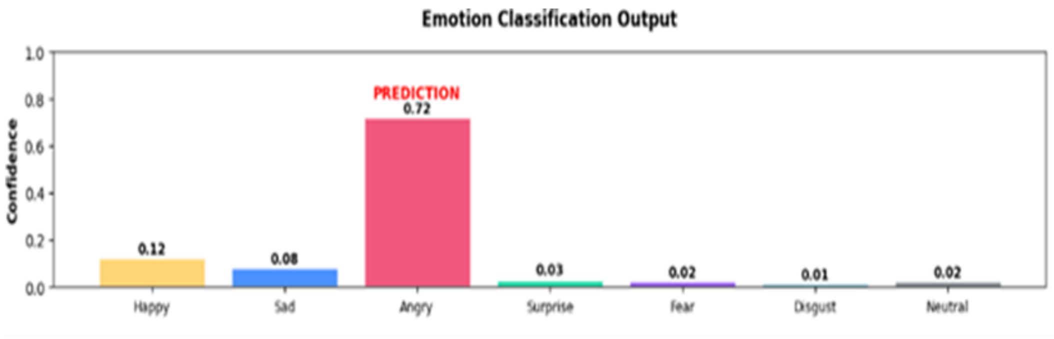


Figure 8: Multimodal complementarity analysis- emotion dimension coverage

6.1. Early Fusion

Early fusion integrates unprocessed or intermediate characteristics across multiple modalities into one representation before the classification.

Mechanism

This method allows the model to acquire cross-modal correlations in the initial phase. Middya et al. [3] established that the early association of acoustic and visual features enhanced accuracy on both the RAVDESS and the SAVEE data sets.

Challenges

Fusion at an early stage necessitates time and frequency correspondence among modalities (Figure 9). Audio and video streams may not have similar sampling rates, which may cause noise on fused representation [15].

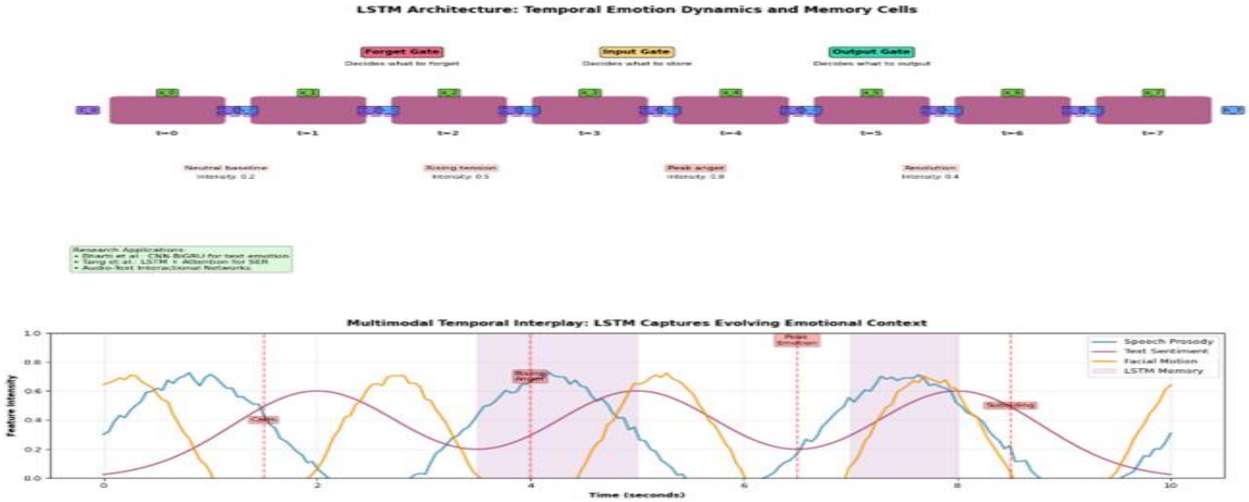


Figure 9: Temporal alignment stream multimodal synchronization of emotion recognition.

6.2. Late Fusion (Decision-Level)

Late fusion models predict separately in each modality (and vote), or weighted average (Figure 8).

Advantages

This is an aggressive approach that is strong. In case a particular modality is not reliable, e.g., poor lighting to read faces, the system can use other modalities. Boitel et al. [11] implemented late fusion using the predictions of DeBERTa, Semi-CNN, and ResNet-50 and achieved better results compared to unimodal baselines.

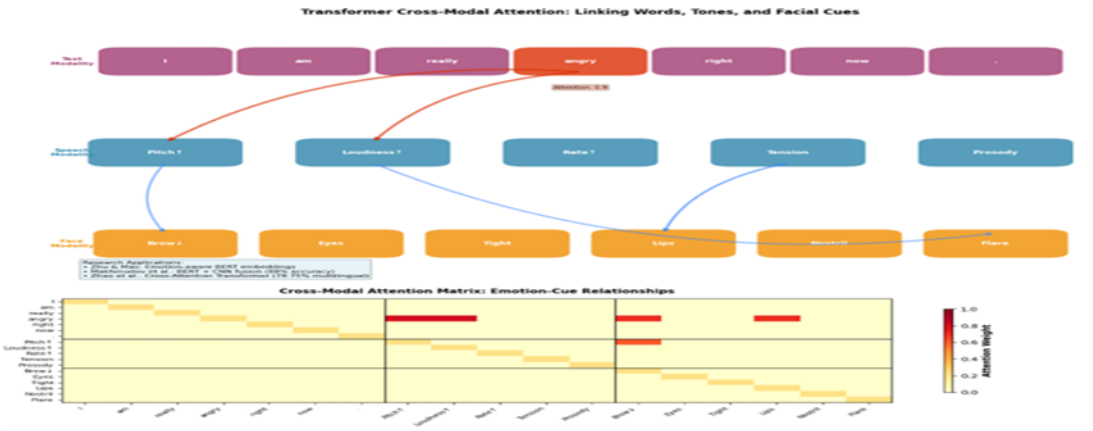

Transformer Cross-Modal Attention: Linking Words, Tones, and Facial Cues
Text Modality
I
am
really
angry
right
now
.
Speech Modality
Pitch ↑
Loudness ↑
Rate ↑
Tension
Prosody
Face Modality
Brow ↓
Eyes
Tight
Lips
Neutral
Flare
Research Applications:
• Zhu & Mao: Emotion-aware BERT embeddings
• Makhmudov et al.: BERT + CNN fusion (88% accuracy)
• Zhao et al.: Cross-Attention Transformer (78.75% multilingual)
Cross-Modal Attention Matrix: Emotion-Cue Relationships
Attention Weight
1.0
0.8
0.6
0.4
0.2
0.0

Figure 10: The last predictions with voting or weighted averaging**.**

6.3. Hybrid and Attention-Based Fusion

The Hybrid fusion strategies utilize attention mechanisms to dynamically estimate the significance of individual modality, as opposed to using fixed weighting schemes [16].

6.3.1. Dynamic Weighting:

In their study, Makhmudov et al. [15] combined BERT-based textual features with the CNN-derived audio features with the help of the attention module readjusting modality focus depending on the emotional intensity obtaining 88.4% accuracy on the CMU-MOSEI dataset.

6.3.2. Transformer-Based Fusion

More complex architectures like MIST and CLIP provide common embedding spaces where modalities can attend to each other. Based on this idea, Chen et al. [4] suggested the LFD-RT architecture that separates the noise of languages and the similarities of emotions, especially in low-resource multilingual scenarios. The resulting confidence of the prediction is shown in Figure 11.

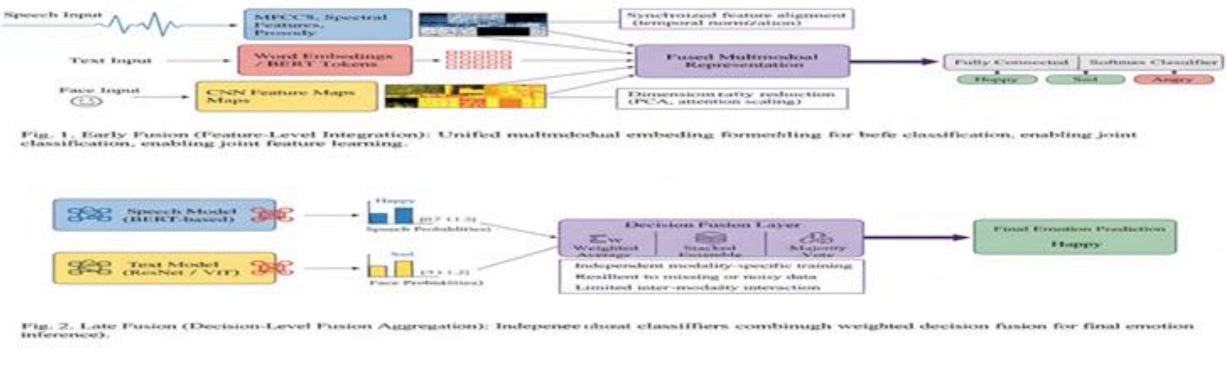

Speech Input
MFCCS, Spectral Features, Prosody
Text Input
Word Embedings / BERT Tokens
Face Input
CNN Feature Maps Maps
Synchroized feature alignment (temporal normalization)
Dimension tatly reduction (PCA, attention scaling)
Fused Multimodal Representation
Fully Connected
Softmax Classifier
Happy
Sad
Angry
Fig. 1. Early Fusion (Feature-Level Integration): Unifed multimdodual embeding formedding for befe classification, enabling joint classification, enabling joint feature learning.
Speech Model (BERT-based)
Text Model (ResNet / ViT)
Happy
Speech Probabilities
Sad
Face Probabilities
Decision Fusion Layer
Σw Weighted Average
Stacked Ensemble
Majority Vote
Independent modality-specific training
Resilient to missing or noisy data
Limited inter-modality interaction
Final Emotion Prediction
Happy
Fig. 2. Late Fusion (Decision-Level Fusion Aggregation): Indepenee dual classifiers combinugh weighted decision fusion for final emotion inference).

Figure 11: Early (feature-level) Vs late (decision-level) fusion strategies

7. Practical Deployment and Latency Considerations

7.1 Computational Cost and Inference Latency

Even though Transformer-based architectures can be used to solve the state-of-the-art problems on benchmark datasets, their implementation is extremely complicated because of significant computational requirements. Although the mechanisms of self-attention can efficiently extract global contextual information, they consume a lot of memory and suffer high amounts of inference latency because of their quadratic complexity with sequence length. In real time systems, including socially interactive robots or emotion sensitive education systems, inference latencies longer than 100200 milliseconds may have adverse effects on system responsiveness and the user experience. Latency constraints are, therefore, an important factor to be considered in implementing such models in a time-sensitive environment.

7.2 Edge Computing and Privacy Considerations

In other applications that require privacy such as personal digital assistants and medical monitoring devices, multimodal emotion recognition would have to be used in real-time at the device to prevent the transmission of sensitive biometric information to the cloud server. This need poses a number of challenges. Mobile and embedded platforms have very limited memory and storage, which limits the size of models and the complexity of architectures. Another important issue is energy efficiency, especially in the case of wearable devices, which need constant monitoring of emotions. Research has shown that Transformer-based models require two to three times the amount of energy to compute an inference compared to optimized convolutional architectures, and that energy-sensitive model development is necessary to ensure sufficient battery life.

The recent method of federated learning has become one of the promising solutions to privacy-preserving model training based on the use of decentralized learning on multiple devices without exchanging raw data. Although this mechanism minimizes the risks of privacy, it brings more problems of heterogeneous hardware capability among devices and communication overhead.

7.3 Real-World Performance Considerations

Laboratory-based performance assessments do not always accurately represent the challenges of the real-world deployment. Even partial occlusion of the face, different illumination conditions, and ambient noise can significantly deteriorate the performance of models as compared to the ones achieved on curated datasets. Moreover, the heterogeneity of hardware deployed into platform (such as processing power, camera resolution, and microphone quality) is another superimposition point that increases complexity. In a bid to guarantee realistic functionality, MER systems consequently require to be developed to perform well under a wide range of environmental factors as well as hardware features.

8. Conclusion

Multimodal emotion recognition Multimodal emotion recognition has developed traditional handcrafted feature-based models to high-level deep learning models. Combining convolutional neural networks with spatial feature extraction, recurrent neural networks with temporal modeling, and Transformer-based architectures based on a global context representation have achieved new performance gains with improvement on benchmark datasets. The latest developments in Transformer-based multimodal fusion with cross-modal attention mechanisms allow the dynamic prioritization of the information that is relevant across the modalities. The increased attention to those environments with low resources and multilingualism only exposes the significance of creating inclusionary and universally applicable emotion recognizing systems. To move forward with MER, computer scientists, psychologists, ethicists as well as domain experts must work interdisciplinarity. By spreading issues pertaining to robustness, interpretability, efficiency, and ethical issues, MER systems may be transformed into both technologically developed and socially responsible systems that provide substantial emotional insight into various applications of the system in the real world.

References

[1] S. N. Atluri and S. Shen, "Global weak forms, weighted residuals, finite elements, boundary elements & local weak forms," in *The Meshless Local Petrov-Galerkin (MLPG) Method,* 1st ed., vol. 1. Henderson, NV, USA: Tech Science Press, pp. 15–64, 2004.

[2] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy, 25(10), 1440.

[3] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A. M., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends, 2(1), 73–79.

[4] Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowledge-Based Systems, 244, 108580.

[5] Chen, L., Guan, S., Huang, X., Wang, W.-J., Xu, C., Guan, Z., & Zhao, W. (2025). Cross-lingual multimodal sentiment analysis for low-resource languages via language family disentanglement and rethinking transfer. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 6513–6522).

[6] Das, R., & Singh, T. D. (2022). A multi-stage multimodal framework for sentiment analysis of Assamese in low resource setting. Expert Systems with Applications, 204, 117575.

[7] Aliyu, Y., Sarlan, A., Danyaro, K. U., Rahman, A. S. B. A., & Abdullahi, M. (2024). Sentiment analysis in low-resource settings: A comprehensive review of approaches, languages, and data sources. IEEE Access, 12, 66883–66909.

[8] Zaidi, S. A. M., Latif, S., & Qadir, J. (2024). Enhancing cross-language multimodal emotion recognition with dual attention transformers. IEEE Open Journal of the Computer Society, 5, 684–693.

[9] Zhao, R., Jiang, X., Yu, F. R., Leung, V. C. M., Wang, T., & Zhang, S. (2025). Leveraging cross-attention transformer and multi-feature fusion for cross-linguistic speech emotion recognition. IEEE Internet of Things Journal, 12(23), 50653–50664.

[10] Zhang, T., & Tan, Z. (2024). Survey of deep emotion recognition in dynamic data using facial, speech and textual cues. Multimedia Tools and Applications, 83(25), 66223–66262.

[11] Bharti, S. K., et al. (2022). Text-based emotion recognition using deep learning approach. Computational Intelligence and Neuroscience, 2022, 2645381.

[12] Boitel, E., Mohasseb, A., & Haig, E. (2025). MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis. Expert Systems with Applications, 270, 126236.

[13] Tang, Y., Hu, Y., He, L., & Huang, H. (2022). A bimodal network based on audio–text-interactional-attention with ArcFace loss for speech emotion recognition. Speech Communication, 143, 21–32.

[14] Makhmudov, F., Kultimuratov, A., & Cho, Y. I. (2024). Enhancing multimodal emotion recognition through attention mechanisms in BERT and CNN architectures. Applied Sciences, 14(10), 4199.

[15] Zhu, Z., & Mao, K. (2023). Knowledge-based BERT word embedding fine-tuning for emotion recognition. Neurocomputing, 552, 126488.

[16] Zhang, X., Mao, R., & Cambria, E. (2024). Multilingual emotion recognition: Discovering the variations of lexical semantics between languages. In 2024 International Joint Conference on Neural Networks (IJCNN) (pp. 1–9).

[17] Makhmudov, F., Kultimuratov, A., & Cho, Y. I. (2024). Enhancing multimodal emotion recognition through attention mechanisms in BERT and CNN architectures. Applied Sciences, 14(10), 4199.

[18] Zhu, Z., & Mao, K. (2023). Knowledge-based BERT word embedding fine-tuning for emotion recognition. Neurocomputing, 552, 126488.

[19] Al-onazi, B. B., et al. (2022). Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion. Applied Sciences, 12(18), 9188.

[20] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

[21] Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53(9–10), 1062–1087.

[22] Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behavior. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005) (Vol. 2, pp. 568–573). IEEE.

[23] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

[24] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125.

[25] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 873–883).

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).

[27] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv.