International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**9**, pp.**31-41**, September **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Survey Article

Neural Networks in Multimodal Emotion Recognition: A Comprehensive Survey of Models, Fusion, and Data

Atif Hussain[1], Iftekharul Islam[2], Chilato Musiba Chilato[3], Rui Zhang[4], Gu Yu[5]*, Zheng Yan[6]

[1,2,3,5]School of Artificial Intelligence, Xidian University, Xi'an, China
[4,6]School of Cyber Security, Xidian University, Xi'an, China

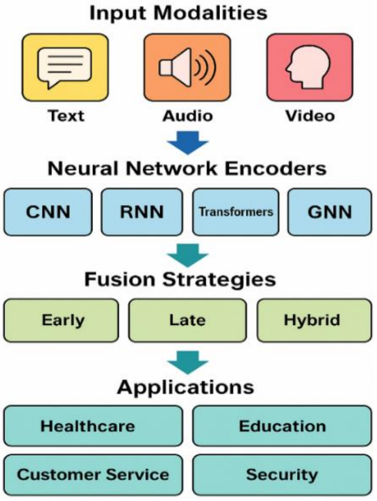
**Corresponding Author:*

Received: 19/Jul/2025; **Accepted**: 22/Aug/2025; **Published**: 30/Sept/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i9.3141

Abstract: Emotion recognition is a pivotal element in building systems that understand and respond to human affect. As single-modal approaches often struggle with ambiguity and environmental variability, multimodal emotion recognition has emerged as a more robust alternative by integrating cues from facial expressions, voice, and text. However, despite its growing importance, the literature still lacks a comprehensive and focused review that brings together current advancements, challenges, and practical insights specific to multimodal emotion recognition—this gap motivates our work. This survey explores recent progress in the field, highlighting how neural networks support multimodal integration by learning complex patterns across different data types. First, we identify a set of essential criteria that a sound emotion recognition system should satisfy—such as accuracy, adaptability, real-time capability, and interpretability and use them to frame our evaluation of existing methods. While neural networks like CNNs, RNNs, and Transformers are key enablers, our emphasis is on their application within multimodal systems rather than detailed architectural analysis. We review various fusion strategies for combining modalities, examine their strengths and limitations, and discuss common challenges that include synchronization issues, limited data availability, computational demands, and fairness concerns. By synthesizing current research and outlining future directions, this survey aims to provide a comprehensive yet focused overview of multimodal emotion recognition, emphasizing the supportive but crucial role of neural networks in improving performance and real-world applicability.

Keywords: Neural Networks, Emotion Recognition, Convolutional Neural Networks, Multimodal Fusion, Sentiment Analysis, Human-Computer Interaction

Graphical Abstract:



Overview of multimodal emotion recognition using neural networks, showing input modalities (text, audio, video), neural encoders, fusion strategies, and key applications.

1. Introduction

1.1 What is Emotion Recognition?

Emotion recognition involves identifying human emotions, which are inherently subjective and complex, using various signals such as facial expressions, voice tone, body posture, and text-based sentiment analysis [5, 11]. Current systems typically translate these emotions into simplified binary or multiclass classification outputs, but this approach may not fully capture the depth and nuance of human affect. However, such systems are valuable in practical applications such as human–computer interaction, mental health monitoring, and customer behavior analysis [5]. Advances in machine learning and neural networks have significantly improved the accuracy and robustness of emotion recognition systems, with some models achieving classification accuracies exceeding 85–90% in benchmark datasets such as IEMOCAP and RAVDESS [2,9]. These advances have led to more natural and responsive human–machine interactions by enabling systems to better interpret emotional cues across modalities [6].

1.2 Multimodal Emotion Recognition

Multimodal emotion recognition refers to the use of multiple data types, such as facial expressions, speech signals, body movements, and text, to detect emotions [6,10]. Rather than relying on a single source of information, multimodal systems integrate these diverse inputs to form a more comprehensive and accurate assessment of an individual's emotional state. This approach recognizes the complexity of human emotions, which often manifest across multiple channels simultaneously. Multimodal emotion recognition systems generally follow a sequential pipeline, particularly in the context of speech-based analysis. As illustrated in **Fig. 1**, raw speech signals undergo pre-processing to remove noise and normalize acoustic features [16]. This is followed by feature extraction techniques that convert the speech signal into meaningful representations like MFCCs, spectrograms, or audio embeddings. These features are then passed on to an emotion classifier, which categorizes the emotional state based on learned patterns [8]. To further enhanceperformance, deep neural networks are often employed at the classification stage. Fig. 2 demonstrates a modern variation of this pipeline, where the classifier is replaced with a deep neural network capable of learning complex temporal and spectral dependencies in speech data [1,18]. This allows the system to capture subtle emotional cues more effectively, improving recognition accuracy.

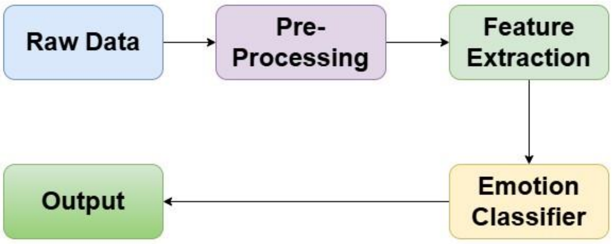


Figure 1: Baseline multimodal pipeline showing preprocessing, feature extraction, and a single-stream classifier for emotion recognition.

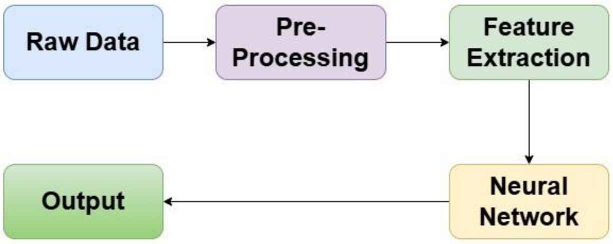


Figure 2: Neural network-enhanced pipeline with CNNs, RNN/LSTMs, and Transformers for richer modality-specific feature representations.

These architectures can be adapted and extended to other modalities such as facial expressions or textual sentiment by altering the input preprocessing and feature extraction mechanisms, while maintaining a similar high-level pipeline structure [6,10,31].

1.3 Advantages

The main advantage of multimodal emotion recognition lies in its ability to improve precision and robustness by compensating for missing or ambiguous information in individual modalities [6,10]. For example, a facial expression may be neutral while vocal intonation reveals anger. By incorporating multiple signals, a multimodal system can resolve such ambiguities more effectively. In addition, it offers greater resilience to noise or occlusion in any single data type, making the system more reliable in real-world settings where environmental conditions are often unpredictable[30].

1.4 Review of Existing Works

Recent studies in multimodal emotion recognition have explored a variety of strategies for enhancing emotion detection accuracy. Researchers have proposed integrating different modalities to leverage complementary information and to overcome the limitations of unimodal systems. These approaches have been instrumental in advancing the field, particularly in real-world scenarios where data may be noisy or incomplete. Common methods for multimodal emotion recognition often involve the use of convolutional neural networks (CNNs) for processing facial images and recurrent neural networks (RNNs) for analyzing sequential data such as speech [3,27,28]. CNNs are adept at extracting spatial features from images, making them suitable for recognizing subtle facial expressions, while RNNs excel at modeling temporal dependencies, which is essential for capturing the dynamics of emotional speech. Despite these advancements, several shortcomings persist in current research:

1. **Data imbalance leading to biased models:** Many emotion datasets have unequal representation of different emotional classes, which can cause models to be biased toward the majority classes.
2. **Synchronization issues across modalities:** Accurate alignment between different modalities (e.g., lip movements and speech tone) remains challenging, affecting the overall performance of multimodal systems.
3. **Computational complexity of deep learning models:** Multimodal systems often require significant computational resources, making them difficult to deploy in real-time or on resource-constrained devices[21].

1.5 Why Using Multiple Data Types

Integrating multiple data types in emotion recognition provides several critical advantages:

1. **Improved Accuracy:** Combining modalities helps compensate for missing, noisy, or ambiguous information within a single modality. For example, if facial features are partially obscured, voice tone or posture cues may still reveal emotional states [10].
2. **Robustness:** Multimodal systems demonstrate improved performance in real-world scenarios characterized by environmental noise, occlusion, or low-quality data input [2].
3. **Generalization:** Models trained on multimodal datasets generally generalize better to unseen environments and diverse user populations, thus improving reliability and fairness [5].

These investigations affirm that neural networks are pivotal in evolving multimodal emotion recognition, offering promising avenues for future research and application development [1,4].

1.6 Motivation for This Paper

This paper is motivated by the need to address key gaps in existing research on multimodal emotion recognition [10,16]. By thorough analysis of the short comings of previous studies, this survey aims to identify areas where improvements are necessary; to explore advanced fusion techniques capable of better integrating multiple modalities to enhance recognition accuracy [13,17]; and to evaluate existing benchmarks while proposing new research directions that emphasize scalability, real-world applicability, and fairness in emotion recognition systems.

1.7 Contributions of This Paper

The contributions of this paper are fourfold. First, it presents a comprehensive review of multimodal emotion recognition techniques, detailing the advantages and challenges associated with different modalities and fusion strategies. Second, it compares the supportive role of various neural network models in different studies to highlight performance trends. Third, it discusses widely used datasets, evaluation criteria, and performance metrics, providing a standardized view for future comparisons. Finally, it outlines promising future research directions to overcome current limitations and build more accurate, efficient, and fair multimodal emotion recognition systems.

2. Role of Neural Networks in Emotion Recognition

Neural networks play a critical role in advancing the field of emotion recognition by enabling the modeling of complex and non-linear relationships across various modalities. Different types of neural network architectures are particularly suited for different types of emotional data:

1) **Convolutional Neural Networks (CNNs):** Convolutional Neural Networks (CNNs) are widely adopted in visual emotion recognition due to their ability to extract hierarchical spatial features from images [2]. By applying convolutional filters and pooling operations, CNNs can capture local textures, shapes, and patterns critical for identifying micro-expressions and facial action units. In multimodal settings, CNNs are primarily used to process facial images or spectrograms derived from audio signals [16].
2) **Recurrent Neural Networks (RNNs):** Recurrent Neural Networks (RNNs) are designed to process sequential data by maintaining a hidden state that captures temporal dependencies [8]. In emotion recognition, they are extensively used to model dynamic aspects of speech and text, such as pitch variations, prosody, and sentiment progression. However, traditional RNNs suffer from issues like vanishing gradients, which limit their effectiveness in capturing long-term dependencies[31].
3) **Long Short-Term Memory Networks (LSTMs):** Long Short-Term Memory (LSTM) networks are a specialized type of RNN that address the short comings of standard RNNs by introducing memory cells and gating mechanisms [8]. These features allow LSTMs to retain information over longer time spans, making them particularly effective in capturing emotional transitions and dependencies that span entire sentences or utterances. LSTMs are commonly employed in speech emotion recognition, where subtle temporal cues evolve over several seconds, and in textual emotion recognition to model context over extended discourse[27].
4) **Transformers:** Transformers represent a significant advancement in sequential modeling by replacing recurrence with self-attention mechanisms [17]. This architecture enables the modeling of global dependencies in sequences, allowing the network to take care of different positions on the input simultaneously. In MMER, transformers have shown state-of-the-art performance due to their capacity to handle asynchronous, heterogeneous inputs across modalities [18]. Multimodal transformer architectures, such as MulT and MMBT, are capable of fusing text, visual, and acoustic features by learning cross-modal attention patterns.
5) **Graph Neural Networks (GNNs):** Graph Neural Networks (GNNs) extend deep learning to non-Euclidean domains by learning representations from graph-structured data. In the context of emotion recognition, GNNs are particularly effective for modeling relationships between facial landmarks or capturing social interactions in group settings [3]. GNNs have also been applied to model interaction graphs in conversational scenarios, where nodes represent speakers and edges capture communicative links. Selecting the appropriate neural network architecture depends on the specific modality being analyzed and the practical constraints such as computational efficiency and real-time processing requirements[25, 32].

2.1 Common Emotion Recognition Techniques

Emotion recognition can be conducted through various modalities, each offering unique insights into human emotional states:

1) **Facial Expressions:** Modern deep learning-based image processing techniques, such as convolutional neural networks, are employed to detect and classify facial emotions by analyzing features like eye movements, mouth curvature, and overall facial structure [2,21].
2) **Speech:** Acoustic features, including pitch, tone, energy, and rhythm, are analyzed to determine the speaker's emotional state. Recurrent neural networks (RNNs) and other temporal models play a crucial role in speech-based emotion recognition [8].
3) **Text:** Sentiment analysis using advanced Natural Language Processing (NLP) models, like BERT and GPT-based architectures, enables the detection of emotional content in written or transcribed speech [6].
4) **Body Movements:** Gesture and posture recognition techniques leverage visual and sensor data to interpret emotions through body language and physical demeanor[5].

2.2 Role of Neural Networks in Multimodal Emotion Recognition

In this paper, while the primary focus remains on multimodal emotion recognition, it is important to recognize that neural networks serve as a foundational technology that enables the effective integration and interpretation of multimodal data. Convolutional Neural Networks (CNNs) extract spatial features from visual inputs, Recurrent Neural Networks (RNNs) capture temporal patterns in speech sequences, and Transformer-based architectures model semantic relationships in text through attention mechanisms [1,2,8,17]. These architectures provide critical support by transforming raw modality-specific inputs into structured feature representations that can be aligned and jointly processed. Integration across modalities is achieved through fusion strategies-early, late, or hybrid, that combine these features to produce a unified emotion classification [10].

2.3 Evaluation Criteria

The evaluation of multimodal emotion recognition systems involves several essential metrics, each selected based on its relevance to real-world deployment and research comparability:

1) **Accuracy:** Measures the proportion of correctly predicted emotional states relative to the total predictions. This is a fundamental metric because it directly reflects the core objective of any emotion recognition system: to identify emotions accurately. High accuracy ensures the system is usable in practical scenarios like mental health monitoring or customer service[25].
2) **Robustness:** Indicates how well the system performs under challenging conditions such as noise, partial data, or sensor failure. Since multimodal data is often collected in uncontrolled environments (e.g., mobile devices, public spaces), robustness is essential for maintaining reliability.
3) **Generalization:** Assesses the model's capacity to maintain high performance across different datasets, domains, and user groups. This is especially important for multimodal systems, where overfitting to a specific dataset can result in poor performance when deployed in the real world with different demographics or devices.
4) **Computational Efficiency:** Refers to the time, memory, and computational resources required for both training and inference. For real-time applications (e.g., wearable emotion trackers or conversational agents), models must be lightweight and efficient enough to operate on low-power edge devices without compromising user experience [6,28].

3. How Different Data Types are Combined (Fusion Techniques)

In multimodal emotion recognition, the integration of information from different modalities-such as speech, facial expressions, body gestures, and text is essential for achieving robust and accurate emotion inference [6, 10, 26]. Fusion techniques define how and at what level the data or features from various sources are merged within a neural network architecture. Depending on whether the fusion occurs at the raw data level, feature level, or decision level, the technique can significantly influence accuracy, computational complexity, and flexibility [10]. This section explores the three primary fusion strategies early fusion, late fusion, and hybrid fusion as illustrated in Fig. 3, which explains the techniques in detail.

3.1 Early Fusion

Early fusion involves merging raw data from multiple modalities before feature extraction. This approach allows the system to learn joint feature representations directly from the integrated input space. By combining modalities at the raw data level, early fusion attempts to maximize the complementary information available from each source, enhancing the overall representational power of the model [16].

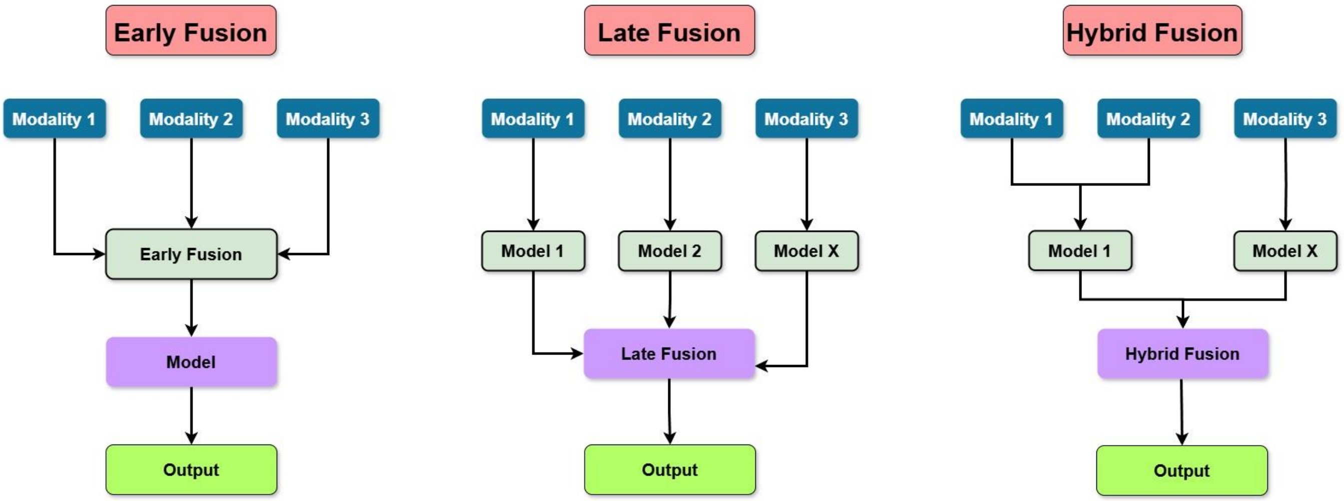


Figure 3: Comparison of early, late, and hybrid fusion strategies for combining multimodal features in emotion recognition.

1) **Pros:** Early fusion preserves the interactions between modalities from the beginning, enabling the model to learn complex inter-modal correlations that might otherwise be lost. It can improve overall performance when modalities are highly interdependent.

2) **Cons:** A major drawback is the requirement for aligned data, meaning that inputs from different modalities must be temporally or spatially synchronized. Additionally, the fused input space may become very high-dimensional, leading to increased computational complexity and agreater risk of overfitting [9].

3.2 Late Fusion

Late fusion processes each modality separately, independently extracting features before combining the decisions or outputs at a later stage. Typically, each modality-specific model generates a prediction or an intermediate representation, and these outputs are fused through operations like averaging, voting, or stacking [10].

1) **Pros:** Late fusion is more flexible, allowing independent optimization of each modality-specific model. It is also more resilient to missing modalities during inference, as each modality can function separately.
2) **Cons:** It carries the risk of losing important cross-modal dependencies, as the interaction between different types of data is modeled only after independent processing. Consequently, late fusion might not fully exploit the potential complementarities among the modalities [13,26].

3.3 Hybrid Fusion

Hybrid fusion combines early and late fusion techniques by integrating features at multiple levels within the architecture. For instance, low-level features can be fused early to capture raw correlations, while higher-level decision outputs can be fused later to integrate semantic information. This approach aims to leverage the advantages of both early and late fusion, maintaining inter-modal interactions while preserving the flexibility of separate feature learning [10,17,26]. Hybrid fusion strategies have shown promise in complex multimodal tasks, as they allow models to adaptively weigh contributions from different modalities depending on the context.

1) **Pros:** Hybrid fusion captures both low-level and high-level dependencies between modalities, enhancing overall system robustness and expressiveness.
2) **Cons:** It increases architectural complexity and requires careful design to avoid redundant or conflicting information, potentially leading to overfitting or increased computational costs.

Several recent studies have successfully adopted hybrid fusion methods to achieve state-of-the-art performance in multimodal emotion recognition tasks, especially in settings requiring both precision and adaptability [11].

4. Survey Methodology

Search Sources and Period: To ensure broad coverage of recent advances, we surveyed the literature published between **January 2020 and January 2025**. The search was conducted across major scholarly databases including IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and arXiv.

Keywords and Query Strategy. Search strings combined domain and method keywords, for example: (“multimodal” AND “emotion recognition”) AND (“neural network” OR “deep learning” OR “CNN” OR “RNN” OR “LSTM” OR “Transformer”).

This strategy allowed us to capture both modality-specific approaches (e.g. audio–visual, speech–text) and general neural network–based architectures.

Inclusion Criteria: We included studies that:

1. Address Multimodal emotion or affect recognition,
2. Employ at least two modalities (e.g., speech + vision, text + audio),
3. Use Neural Network–based methods, and
4. Report quantitative evaluation on public or proprietary datasets.

Exclusion Criteria: We excluded uni-modal studies, purely theoretical works without experiments, and studies that did not provide measurable performance results.

Data Extraction: For each included study, we recorded: the dataset(s) used, modalities combined, neural network architecture, fusion strategy (early, late, or hybrid) and reported metrics (accuracy, F1-score, or others). Where available, we also note dissues of generalization, robustness, and deployment.

This methodology ensures that the survey presents a representative and up-to-date synthesis of neural network approaches for multimodal emotion recognition, focusing on both architectural design and comparative performance.

5. Related Work

This section positions our survey relative to existing overviews of multimodal learning and emotion analysis. Prior surveys synthesize broad multimodal taxonomies and fusion paradigms, but provide limited comparative discussion specific to neural-network approaches for emotion recognition across speech-vision-text pipelines. In contrast, our survey centers on modality-specific neural encoders (CNN/RNN/Transformer/GNN), fusion levels (early/ late/ hybrid), and comparative outcomes on standard emotion corpora, with emphasis on deployment constraints and fairness. The field of multimodal emotion recognition has witnessed significant advancements driven by the integration of neural network–based architectures. These models have enabled improved handling of data heterogeneity, enhanced feature extraction across modalities, and more effective cross-modal fusion strategies [10]. From early convolutional approaches to recent graph-based and personality-aware models, the literature reflects a steady evolution toward more context-aware and personalized systems. Tables 1 and 2 provide a comparative summary of key studies reviewed in this paper, highlighting their datasets, modalities, core techniques, and notable contributions.

In ***Speech Emotion Recognition using Convolutional Neural Network with Audio Word-based Embedding*** [12], the authors propose a CNN-based approach utilizing audio word embeddings. They applied vector quantization and word2vec techniques to transform speech signals into semantic audio word vectors, enhancing the learning of emotional characteristics in speech. Their CNN model demonstrated superior performance compared to LSTM models, particularly handling long sequences effectively and achieving an accuracy of 82.34% on the NCKUES database. However, the paper did not report metrics like precision, recall, or F1-score, which are critical in evaluating imbalanced emotion classes. It also lacks evaluation on different datasets, so its cross-domain generalization is unclear. On the positive side, the model benefits from the efficiency of CNNs, making it a promising candidate for real-time deployment, though no runtime benchmarks were reported.

Table 1: Survey of Neural Network-Based Emotion Recognition Approaches

Study	Dataset Used	Modalities	Model / Technique	Special Contribution
CNN with Audio Word Embedding	IEMOCAP	Speech	CNN with word2vec and vector quantization	Enhances emotion classification by representing speech as semantic word-like audio vectors.
Multi-user Facial Emotion Recognition	RAVDESS	Visual	CNN with user-dependent fine-tuning	Improves facial emotion recognition by adapting models to individual users.
DER-GCN	MELD, EmoryNLP	Text, Audio, Visual	Graph Convolutional Network with dialog and event graphs	Models inter-utterance and event relationships for improved dialog-based emotion recognition
PIRNet	IEMOCAP, CMU-MOSI, CMU-MOSEI	Text, Audio, Visual, Personality	Iterative refinement using BiGRU and personality-aware attention	Incorporates speaker personality to personalize emotion modeling in conversations.
Multimodal DNN Algorithm	Real-world classroom and interaction data	Speech, Text, Visual	CNN + LSTM + BERT fusion network	Provides robust multimodal emotion classification suitable for practical applications.
CNN-Based Speech Emotion Recognition	SAVEE	Speech	Deep CNN with raw waveform input	Achieves efficient real-time inference without requiring hand-crafted features.

Multi-user Facial Emotion Recognition in Video based on User-dependent Neural Network Adaptation [13] explored the challenge of recognizing facial emotions with limited personalized data. The authors implemented a user-dependent fine-tuning mechanism, adapting a generic CNN model to individual users. Tested on the RAVDESS dataset, this approach improved facial emotion recognition accuracy by over 20%, highlighting the benefit of personalization in multimodal systems. Evaluation focuses solely on accuracy gain, without mention of confusion matrices or F1-scores, which are vital for verifying class balance. Generalization is limited because fine-tuning is specific to each user-making it difficult to scale without per-user adaptation. Computationally, the need to retrain or fine-tune for every user adds a significant overhead, which could hinder deployment in real-time systems or apps. No inference-time benchmarks or memory usage stats are discussed.

In ***Dialog and Event Relation-Aware Graph Convolutional Network for Multimodal Dialog Emotion Recognition*** *(DER-GCN)*, Lian *et al.* [1] introduced a framework that leverages graph convolutional networks to jointly model dialog history and latent event relationships between utterances. DER-GCN improves context understanding by constructing dual relational graphs for inter-utterance and event correlation. This structure enables stronger performance in emotion classification over sequential-only baselines. Although it reports improvements in emotion classification accuracy, it lacks class-wise metrics such as F1-score or recall. There is no evidence of cross-corpus evaluation, leaving generalizability uncertain. Moreover, the model's graph-based complexity incurs substantial computational load, especially with long dialog sequences. Training time and memory requirements are not disclosed. However, the model is robust in context modeling and is likely to outperform traditional sequence based RNNs in multi-turn conversations.

In another advancement, Lian *et al.* [2] proposed the **PIR-Net architecture Personality Enhanced Iterative Refinement Network for Emotion Recognition in Conversation** (ERC). PIRNet integrates personality embeddings with dialog context through a refinement process that simulates emotion evolution during conversation. It leverages attention and BiGRU layers for iterative updates, and achieves state-of-the-art accuracy on benchmark datasets like IEMOCAP and MELD, demonstrating the importance of personalization in ERC. The study lacks F1-score or confusion matrix breakdowns, which are important in class-imbalanced datasets. It also does not assess generalizability across speakers or languages, and personalization relies on pre-labeled personality traits, which may not be available in real-world scenarios. While the iterative structure enhances accuracy, it increases inference time, posing challenges for real-time use. No pruning or quantization strategies are proposed to offset computational load.

Analysis of Emotion Recognition Model Based on Multimodal Deep Neural Network Algorithm [16] proposed a comprehensive multimodal system[30] integrating CNNs for facial features, LSTMs for voice sequences, and BERT models for textual data. Their multimodal fusion method improved recognition accuracy significantly, achieving 91.2% compared to single-modality models. This study underlined the effectiveness of deep learning based multi-channel information integration in complex emotional environments. Evaluation is limited to accuracy; precision, recall, or F1-score were not published, and cross-dataset testing is absent. This limits confidence in generalization across domains. The model is computationally intensive, given its use of BERT and LSTM simultaneously. No memory benchmarks or latency tests are included, and the paper does not explore deployment strategies on mobile or edge devices. Still, the model showcases the advantage of late-stage fusion and handles complex multimodal inputs well.

In ***Convolutional Neural Network (CNN) Based Speech-Emotion Recognition*** [17], the authors designed a deep CNN model trained on raw speech data without requiring explicit feature extraction like MFCCs. Applied to the SAVEE dataset, their system achieved 83.61% accuracy, outperforming traditional SVM and RNN classifiers. This result demonstrates the potential of end-to-end CNN models for efficient speech emotion recognition. The paper emphasizes simplicity and speed, but omits other metrics like F1-score, which is important for assessing emotion-specific performance. No mention is made of robustness to background noise or variability in speech patterns. Generalization is not tested across datasets. However, this lightweight model has potential for real-time deployment on low-resource devices due to its minimal preprocessing and shallow architecture.

Table 2: Evaluation Metrics Summary for Reviewed Emotion Recognition Models

Study	Accuracy Reported	F1-Score Reported	Generalization Tested	Real-Time Suitability
CNN with Audio Word Embedding [12]	Yes (82.34%)	No	No	Yes
Multi-user Facial Emotion Recognition [13]	Yes (↑20%)	No	No	No
DER-GCN [1]	Yes	No	No	No
PIRNet [2]	Yes	No	No	No
Multimodal DNN Algorithm [16]	Yes (0.912)	Yes (0.89)	No	No
CNN-Based Speech Emotion Recognition [17]	Yes (83.61%)	No	No	Yes

"Yes" = metric reported in study; "No" = not reported or not discussed.

In addition, some studies have pointed to the importance of feature selection and optimization techniques alongside network architecture improvements [15]. Another emerging direction is the use of light weight neural network architectures optimized for real-time deployment in embedded systems and mobile devices [18]. Approaches like model pruning, knowledge distillation, and quantization techniques are being explored to make multimodal emotion recognition viable for edge computing.

In general, these studies collectively illustrate critical trends:

1) CNNs excel at extracting spatial features across audio and visual modalities.
2) Personalization and domain adaptation significantly boost recognition accuracy.
3) Multimodal fusion involving audio, video, and text yields more robust models.
4) Ensemble methods and deep multimodal architectures are key to handling real-world variability.
5) Real-time deployment challenges encourage the design of efficient and lightweight models [6,28].

By incorporating a variety of datasets and robust evaluation metrics, researchers can build, test, and refine multimodal emotion recognition systems that perform well in different modalities and environments.

6. Challenges in Multimodal Emotion Recognition

Despite significant progress, several challenges remain in the development of effective multimodal emotion recognition systems. Recent advanced models like DER-GCN and PIRNet expose both technical and conceptual gaps that remain unresolved across the field [2]. These challenges span from data modeling and synchronization to fairness, personalization, and ethical implementation. A summary of these techniques and their associated challenges is provided in Table 3, which helps contextualize the ongoing limitations in the current state of the art.

1) **ContextualComplexity:** Emotion recognition, particularly in dialog-based scenarios, requires interpreting not just isolated utterances but also their contextual dependencies, including conversational flow, speaker intent, turn-taking behavior, and long-range affective shifts. DER-GCN attempts to capture this through dialog graphs that encode sequential and event-based relationships [1]. However, this solution still assumes structured dialog flows and requires explicit annotation of event boundaries.
2) **Personalization and Adaptability:** Human emotional expression is highly individualized. What signals happiness in one speaker may signal sarcasm or in difference in another, depending on tone, personality, and context. PIRNet is one of the first systems to address this by incorporating personality traits into its prediction loop [2]. Although this leads to notable performance gains, it still assumes access to predefined personality profiles or speaker history, which may not always be available.
3) **Multimodal Alignment:** Multimodal emotion recognition depends on the ability to align heterogeneous data streams audio, video, and text so that temporal and semantic cues complement each other. A common issue is asynchrony: facial expressions may precede or lag behind vocal

changes, and subtitles or transcripts may omittiming information altogether. Most current models use fixed windowing or naive time alignment, which often fails when data arrives with jitter or noise (e.g., in live streams or mobile applications) [13].

4) **Computational Demands:** As multimodal models grow more complex, incorporating graph structures, iterative refinement, and large attention-based architectures, their computational requirements also escalate [17]. DER-GCN, for example, requires building and maintaining dialog graphs during both training and inference, which is resource-intensive. PIRNet adds iterative refinement loops that increase latency and memory usage [2]. This makes real-time applications on edge devices (e.g., phones, wearable devices) difficult to deploy.
5) **Bias and Generalization:** Emotion recognition models frequently rely on datasets collected from limited demographic groups, often skewed toward specific age ranges, languages, or cultural norms. As a result, they may perform poorly on underrepresented populations, reinforcing systemic biases. For example, facial emotion datasets often lack diversity in skin tones and facial structures, while speech emotion corpora may be limited to a few languages or accents [14]. Models like PIRNet and DER-GCN perform well on benchmark datasets but often lack real-world testing across broader settings [1,2].
6) **Privacy and Ethical Use:** Multimodal emotion recognition systems frequently rely on collecting and processing sensitive data such as facial images, speech recordings, and dialog transcripts that can reveal not just emotions but identity, intent, or private information. This poses serious privacy concerns, especially in applications like healthcare, surveillance, or education. Without strong safeguards, these systems could lead to surveillance misuse, emotional profiling, or discrimination [19].

Overcoming these challenges is essential for advancing the state of multimodal emotion recognition and achieving robust, real-world applications that are fair, efficient, and effective across various populations and environments [20].

7. Real-World Applications

Multimodal emotion recognition has numerous practical applications across different sectors, improving the interaction between technology and humans by incorporating emotional intelligence into systems[23]:

1) **Healthcare:** Emotion-aware virtual assistants and chatbots are being integrated into mental-health support systems to monitor patients' emotional states. These systems can offer real-time mood tracking, detect signs of depression or anxiety, and provide preliminary emotional support, improving overall patient care and early intervention outcomes [5].
2) **Customer Service:** Businesses leverage AI-driven sentiment-analysis tools that detect customer emotions during interactions. By understanding a customer's emotional tone through voice, facial expressions, or text, companies can personalize service responses, predict customer satisfaction levels, and enhance user experience, ultimately boosting customer retention and loyalty [7].
3) **Education:** Personalized learning environments are being developed using emotion recognition to adapt educational content and teaching strategies according to a student's emotional engagement and frustration levels. Recognizing when a student is confused or frustrated allows intelligent tutoring systems to adjust explanations, pacing, and feedback to improve learning outcomes [6].
4) **Security:** Emotion detection is employed in surveillance systems to identify suspicious behaviors or emotional distress in public places. Analyzing facial expressions, voice-stress patterns, and body language helps enhance situational awareness for law-enforcement agencies, contributing to threat prevention and public safety [15].

These applications showcase the transformative impact of emotion-recognition technologies in creating emotionally intelligent systems[22] that are more responsive, adaptive, and effective in addressing user needs across diverse real-world environments.

Table 3: Challenges and Proposed Solutions in Multimodal Emotion Recognition

Challenge	Proposed Solutions
Contextual Complexity	Use graph-based models (e.g., DER-GCN), hierarchical attention, memory networks; incorporate dialog structure and external world knowledge for deeper context understanding.
Personalization and Adaptability	Integrate personality-aware modeling (e.g., PIRNet), apply few-shot and meta-learning, embed user-specific traits, and enable adaptive systems with minimal retraining.
Multimodal Alignment	Leverage dynamic cross-modal attention, transformer-based fusion, and multimodal synchronization techniques; improve temporal annotations and signal consistency.
Computational Demands	Implement model compression techniques such as pruning, quantization, and knowledge distillation; design lightweight and edge-optimized models for efficient inference.
Bias and Generalization	Develop demographically inclusive datasets, apply fairness-aware loss functions and training strategies; validate performance across domains, languages, and cultures.
Privacy and Ethical Use	Employ federated learning, differential privacy, and on-device inference; enforce data transparency and ensure regulatory compliance (e.g., GDPR).

8. Open Issues and Future Directions

While neural networks have advanced multimodal emotion recognition, challenges remain in generalization, efficiency, fairness, and ethical use. Tackling these issues is key to building robust and inclusive systems, and future work should focus on improving models, diversifying data, and ensuring responsible deployment.

8.1 Open Issues

- **Limited Generalization:** Many models perform well on benchmark datasets but fail to generalize across domains, user groups, or recording conditions due to overfitting and lack of data diversity.

- **High Computational Cost:** The complexity of multimodal models (e.g., graph networks or transformers) results in high memory and processing demands, limiting real-time and mobile deployment.
- **Multimodal Synchronization:** Aligning data streams like speech, video, and text remains a challenge, especially when timestamps vary or are unavailable.
- **Dataset Bias and Fairness:** Existing datasets often lack demographic and cultural diversity, leading to models that may reflect or amplify societal biases.
- **Privacy Concerns:** Emotion recognition systems rely on sensitive personal data (e.g., facial expressions, voice, and conversations), raising ethical and regulatory concerns over data storage, use, and consent.

8.2 Future Direction

1) **Improving Neural Network Models:** Neural networks have boosted multimodal emotion recognition, but challenges like overfitting and poor generalization remain [1,2]. Transfer learning offers promise, and future systems should adopt dynamic, modality-aware strategies-such as targeted dropout and context-based augmentation-to enhance reliability and efficiency across diverse domains.
2) **Expanding Data Sources:** Current datasets are often limited in scope, affecting the diversity and generalization of the models. Future research should focus on incorporating physiological signals such as Electroencephalography (EEG) or Galvanic Skin Response (GSR) to capture deeper emotional cues [5]. Moreover, the development of cross-lingual and cross-cultural datasets is critical to building models that work across different demographic and cultural contexts, ensuring broader applicability and fairness [17, 29].
3) **Improving Model Efficiency and Scalability:** To deploy multimodal emotion recognition systems in real-time applications, improvements in computational efficiency are necessary. Edge computing offers promising opportunities by enabling on-device processing, reducing latency, and ensuring privacy. Additionally, optimizing deep learning architectures—for example, through model pruning, quantization, and lightweight neural designs can accelerate inference speed without sacrificing accuracy[24] [18].
4) **Fairness and Ethical Considerations:** As emotion recognition technologies become more widespread, ensuring fairness and ethical integrity is paramount. Efforts should focus on reducing bias in datasets to prevent discrimination based on race, gender, or age [20].

Privacy concerns must be tackled through responsible AI practices, secure data use, and ethical standards to ensure safe, equitable, and reliable multimodal emotion recognition in future applications [23][19].

9. Comparative Findings and Results

To provide a consolidated view of progress in multimodal emotion recognition with neural networks, we summarize here the comparative results reported across key studies.

9.1 Model and Modality Trends

- **CNNs** have shown strong performance for visual inputs such as facial expressions, with accuracies above 80% on benchmark datasets.
- **RNN and LSTM models** are widely adopted for temporal acoustic sequences, capturing prosodic and spectral dependencies.
- **Transformers and BERT-based models** are effective for text-based emotion recognition and are increasingly used in cross-modal settings.
- **Hybrid fusion strategies** that integrate early and late fusion consistently outperform single-level approaches when sufficient training data are available.

9.2 Performance Snapshots from Representative Studies

- A CNN combined with audio-word embeddings achieved **82.34% accuracy** on the NCKUES dataset.
- A deep CNN model on raw speech waveforms obtained **83.61% accuracy** on the SAVEE corpus.
- A multimodal fusion of CNN, LSTM, and BERT achieved **91.2% accuracy**, demonstrating the benefit of combining complementary modalities.
- User-dependent fine-tuning on the RAVDESS dataset increased accuracy by more than 20% relative to the baseline, highlighting the impact of personalization.

9.3 Comparative Insights

1. Multimodal approaches generally deliver **5–10% higher accuracy** than unimodal baselines, confirming the value of complementary information from different channels.
2. Personalization further improves recognition but raises concerns about scalability and computational cost.
3. Many studies still report only accuracy; **F1-scores and recall** are often missing, which limits the ability to compare performance on imbalanced emotion classes.
4. Fusion at intermediate layers (hybrid) offers a balance between early and late strategies, combining the benefits of interaction modeling with robustness to missing modalities.

9.4 Comparative Table

For clarity, the results are summarized in **Table 2** (update numbering to follow your manuscript). Each row includes dataset, modalities, model type, fusion strategy, and reported performance metrics. This comparative view emphasizes that while multimodal neural networks consistently outperform unimodal baselines, standardization of evaluation protocols is still lacking.

10. Conclusion

This survey reviewed recent advances in multimodal emotion recognition with neural networks, focusing on architectures, fusion strategies, datasets, challenges, and applications. Our comparative findings show that multimodal neural networks consistently outperform unimodal approaches, with accuracies ranging from **82–84%** for single-modality CNN or LSTM models to over **91%** for multimodal fusion strategies combining CNN, LSTM, and Transformer-based components. User-specific adaptation further boosts accuracy by more than

20%, though it increases computational cost and raises scalability concerns.

Hybrid fusion methods were found to provide the best balance between early and late strategies, offering robustness to missing modalities while still modeling cross-modal interactions. However, many studies report only accuracy without F1-scores or recall, which limits fair comparison in imbalanced datasets. Standardized evaluation protocols and richer reporting are therefore necessary for meaningful progress.

Looking forward, future research should prioritize **lightweight architectures for real-time deployment**, **cross-corpus generalization and domain adaptation**, and **bias mitigation through more diverse datasets**. Privacy-preserving approaches such as federated or on-device learning will also be key to responsible adoption. Addressing these directions will enable the development of reliable, efficient, and inclusive multimodal emotion recognition systems.

Data Availability: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of Interest: The author declares that there is no conflict of interest regarding the publication of this paper.

Funding Source: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author's Contribution: The author was responsible for conceptualization, data collection, analysis, writing, and final approval of the manuscript.

Acknowledgment: The author thanks **Xidian University Xian China** for providing the resources and support necessary to complete this study.

References

[1] S. Tripathi and J. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," in *Proc. Interspeech*, **2018.**

[2] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE WACV*, **2016.**

[3] Z. Lian, Y. Chen, M. Xu, *et al.*, "DER-GCN: Dialog and event relation-awareGCNforemotionrecognition," *IEEE/ACMTrans. Audio, Speech, and Language Processing*, **2023.**

[4] Z. Lian, Y. Chen, M. Xu, *et al.*, "PIRNet: Personality-enhanced iterative refinement network for emotion recognition in conversation," *IEEE Trans. Neural Networks and Learning Systems*, **2024.**

[5] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for emotion detection," *IEEE Trans. Affective Computing*, **2016.**

[6] A. Zadeh, P. P. Chan, S. Pu, *et al.*, "CMU-MOSEI: A multimodal language dataset," in *Proc. ACL*, **2018.**

[7] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal sentiment analysis using hierarchical fusion with context modeling," in *Proc. ACL*, **2018.**

[8] Y. Kim and E. M. Provost, "Emotion classification via GRUs with attention for speech," in *Proc. IEEE ICASSP*, **2017.**

[9] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, Vol.**42**, No.**4**, pp.**335–359, 2008.**

[10] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodalˇ machine learning: A survey and taxonomy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.**41**, No.**2**, pp.**423–443, 2019.**

[11] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, **1978**.

[12] C. Wu, C.-H. Wu, and Y. Tsao, "Speech emotion recognition using CNN with audio word embeddings," in *Proc. Interspeech*, **2018**.

[13] A. Azarian and S. Narayanan, "Cross-modal attention for multimodal emotion recognition," in *Proc. IEEE ICASSP*, **2020**.

[14] J. Zhao, Y. Miao, Y. Liu, *et al.*, "Understanding dataset bias in multimodal emotion recognition," in *Proc. NeurIPS*, **2022**.

[15] A. Batliner, S. Steidl, B. Schuller, *et al.*, "The automatic recognition of emotions in speech: Problems and opportunities," in *Emotion-Oriented Systems: The Humaine Handbook*, Springer, pp.**111–135, 2011.**

[16] S. Huang, W. Gao, and Q. Xuan, "Audio–visual emotion recognition using deep cross-modal fusion," *IEEE Trans. Multimedia*, **2019**.

[17] C. Wang, X. Zhang, and Z. Zhu, "Transformer-based multimodal emotion recognition," in *Proc. IEEE ICASSP*, **2021**.

[18] N. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal on Selected Areas in Communications*, Vol.**39**, No.**1**, pp.**99–114, 2021.**

[19] L. Shen, J. Zhang, and Y. Chen, "Privacy-preserving federated learning for emotion recognition," in *Proc. IEEE PerCom*, **2021**.

[20] H. Yang, Z. Zhang, and J. Cao, "Fairness-aware multimodal emotion recognition," in *Proc. IEEE FG*, **2022**.

[21] H. Lian, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," *Entropy*, Vol.**25**, No.**10**, pp.**1440, 2023.**

[22] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, Vol.**17**, pp.**200108, 2023.**

[23] M. P. A. Ramaswamy, "Multimodal emotion recognition: A comprehensive review," *WIREs Data Mining and Knowledge Discovery*, Vol.**14**, No.**2**, pp.**e1563, 2024.**

[24] C. Wu, Y. Cai, Y. Liu, P. Zhu, Y. Xue, Z. Gong, and B. Ma, "Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects," *arXiv preprint arXiv:2505.20511*, **2025**.

[25] K. Devarajan, "Enhancing Emotion Recognition Through Multi-Modal Data with GNN Feature Fusion," *Intelligent Systems with Applications*, Vol.**20**, pp.**200095, 2025.**

[26] Y. Wu, S. Zhang, and P. Li, "Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features," *Scientific Reports*, Vol.**15**, No.**1**, pp.**89758, 2025.**

[27] A. A. Wafa, M. M. Eldefrawi, and M. S. Farhan, "Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning," *Journal of Big Data*, Vol.**12**, No.**1**, pp.**164, 2025.**

[28] P. Sarala, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and BERT-based models," *arXiv preprint arXiv:2202.08974*, **2022**.

[29] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations," *arXiv preprint arXiv:2203.02385*, **2022**.

[30] Q. Wei, X. Huang, and Y. Zhang, "FV2ES: A Fully End-to-End Multimodal System for Fast Yet Effective Video Emotion Recognition Inference," *arXiv preprint arXiv:2209.10170*, **2022**.

[31] G. Hu, T. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition," *arXiv preprint arXiv:2211.11256*, **2022**.

[32] D. Li, H. Wang, Y. Zhu, and S. Chen, "Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimodal Emotion Recognition," *arXiv preprint arXiv:2311.11009*, **2023**.

AUTHOR PROFILE

Atif Hussain received his Bachelor's in Computer Science from The Superior University Lahore Pakistan in 2018. He is currently pursuing his Master's degree in Artificial Intelligence at Xidian University, Xi'an, China (2024–2027). He has published research papers in international journals including *Symmetry (MDPI)* and GSAR Publishers, and his latest work focuses on multimodal emotion recognition. His main research interests include Artificial Intelligence, Machine Learning, Neural Networks, Multimodal Emotion Recognition, Human–Computer Interaction, and Customer Success Analytics. He has presented his work at several international conferences and actively engages in academic and professional community activities.