American Journal of Artificial Intelligence
2026, Vol. 10, No. 2, pp. 198–208
https://doi.org/10.11648/j.ajai.20261002.13

SciencePG
Science Publishing Group

Review Article

A Systematic Review of Emotion Recognition: From Unimodal Signals to Multimodal Integration

Bilkisu Muhammad Bashir[1,*], Zayyanu Yunusa[2]

[1]Department of Computer Science, Skyline University Kano, Kano, Nigeria
[2]Department of Computer Science, Modibbo Adama University Yola, Yola, Nigeria

Abstract

Emotion recognition is a core component of affective computing, enabling intelligent systems to interpret human emotional states across critical applications such as healthcare, online education, and human–computer interaction. Early unimodal approaches relying solely on facial expressions, speech, or text have proven insufficient due to noise, cultural variability, and signal ambiguity, prompting a decisive shift toward multimodal integration. This systematic review, conducted following the PRISMA framework, examines this transition by analyzing 89 peer-reviewed studies selected from an initial pool of 160. The objective is to synthesize current methodologies, compare performance across modalities, and identify persistent technical and ethical barriers. Our findings reveal that multimodal systems, which fuse visual, acoustic, linguistic, and physiological signals, consistently outperform unimodal counterparts, achieving accuracy levels above 85% on benchmark datasets. Deep learning architectures particularly convolutional networks for spatial features, recurrent networks for temporal dependencies, and transformer-based models enhanced with attention mechanisms dominate the field, enabling effective dynamic weighting and fusion of heterogeneous data streams. Despite these advances, several challenges impede real-world deployment. Cross-subject and cross-session variability degrades generalizability, while data scarcity and the lack of large-scale, annotated multimodal corpora constrain model training. Computational complexity, especially in transformer-based fusion, limits edge-device feasibility, and ethical concerns surrounding privacy, demographic bias, and model interpretability remain unresolved. Future research must prioritize scalable and lightweight architectures, inclusive and culturally diverse dataset curation, and explainable AI frameworks that build user trust. Ultimately, transitioning these systems from laboratory prototypes to ethically sound, practical applications will require close interdisciplinary collaboration among computer scientists, psychologists, and ethicists, ensuring that emotion recognition technologies are not only accurate but also fair, transparent, and accessible across diverse real-world settings.

Keywords

Multimodal Emotion Recognition, Unimodal Emotion Recognition, Deep Learning, Affective Computing, PRISMA, Transformer Based Architecture

[*]Correspondence: Bilkisu Muhammad Bashir (bilkisubashirbilhaq@gmail.com)

Received: 30 June 2026; **Accepted:** 9 July 2026; **Published:** 11 August 2026

Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Emotion recognition has emerged as a cornerstone of affective computing, with profound implications for human-computer interaction, mental health assessment, education, healthcare, and social robotics. Emotions fundamentally influence how individuals make decisions, plan, reason, and interact with their environment, making automated human emotion recognition (AHER) a critical research priority across multiple disciplines [18]. The ability to automatically detect and interpret emotional states from behavioral and physiological signals has transformative potential for developing empathetic artificial intelligence systems that can respond appropriately to human affective needs.

The proliferation of social media platforms, wearable sensors, mobile devices, and deep learning technologies has catalyzed significant advances in emotion recognition research over the past decade [4]. Social media users generate vast amounts of emotionally charged textual content, images, and videos daily, creating unprecedented opportunities for large-scale affective analysis. Simultaneously, the widespread adoption of smartwatches, fitness trackers, and other wearable devices has enabled continuous, non-invasive monitoring of physiological signals including heart rate, galvanic skin response, and skin temperature all of which carry valuable information about emotional states [11]. These technological developments have been complemented by breakthroughs in deep learning architectures, particularly convolutional neural networks, recurrent neural networks, transformers, and graph neural networks, which have dramatically improved the accuracy and robustness of emotion classification systems [12].

Despite these advances, fundamental challenges persist that limit the real-world deployment of emotion recognition systems. Cross-subject generalization remains a primary obstacle, as models trained on one population often fail to generalize to new individuals due to anatomical, cultural, and experiential differences in emotional expression [2]. Data scarcity particularly for negative emotions such as fear, disgust, and sadness constrains model training and leads to biased performance favoring neutral and positive emotions [10]. The inherent complexity and subjectivity of human emotional experience, which varies across cultures, contexts, and developmental stages, further complicates the development of universally applicable recognition systems [13].

This systematic review aims to provide a comprehensive analysis of current emotion recognition methodologies, focusing on the transition from unimodal to multimodal approaches. The review addresses three primary research questions:

1) What modalities and fusion techniques are currently employed in emotion recognition systems?
2) What are the relative performances of unimodal versus multimodal approaches?
3) What are the primary challenges and future directions in this field?

2. Literature Review

The reviewed literature demonstrates a clear evolutionary trajectory in emotion recognition research, transitioning from unimodal approaches (relying solely on facial expressions, speech, text, or EEG signals) to sophisticated multimodal fusion frameworks that consistently achieve superior performance [18]. Key findings indicate that multimodal systems integrating two or more modalities particularly combinations of facial, vocal, textual, and physiological signals regularly achieve classification accuracies exceeding 85-95% on benchmark datasets such as IEMOCAP, SEED, and MELD, outperforming unimodal baselines by substantial m argins [6]. Transformer-based architectures with attention mechanisms have emerged as the dominant paradigm for cross-modal integration, with models like COLIN and M-fusHER achieving state-of-the-art results by dynamically weighting modality contributions and capturing complex inter-modal dependencies [6].

Despite these advances, several critical challenges remain unresolved. Cross-subject variability continues to limit real-world deployment, with EEG-based systems showing dramatic performance degradation from 97.7% on controlled datasets (SEED) to 51.3% on more complex multi-class datasets (MPED) when moving to real-world scenarios [2]. Label imbalance, data scarcity for negative emotions (fear, disgust), and the computational demands of deep learning models pose additional barriers to practical implementation [4]. Emerging research directions emphasize explainable AI for transparent decision-making [16], domain adaptation techniques for rapid personalization [1], and ethical frameworks addressing privacy, consent, and bias particularly critical as emotion recognition systems are deployed in sensitive domains including education, healthcare, and gender-based violence detection [11]. The integration of affective computing with educational technology [7], elderly care robotics, and mental health assessment represents a promising frontier, with preliminary studies demonstrating significant behavioral improvements and therapeutic outcomes [1]. Future progress will depend on developing large-scale, culturally diverse multimodal datasets, advancing few-shot and zero-shot learning techniques, and establishing standardized evaluation protocols that enable meaningful cross-study comparisons [9].

2.1. The Evolution from Unimodal to Multimodal Approaches

Early emotion recognition research focused predominantly on unimodal approaches, analyzing isolated channels of emotional information. Facial expression recognition systems examined muscle movements and facial action units [13]. Speech emotion recognition systems extracted prosodic and spectral features from vocal signals [6]. Text-based detectors analyzed linguistic content using lexicon-based methods or machine learning classifiers [9]. EEG-based systems measured neural correlates of emotional processing through scalp electrodes [5]. While each unimodal approach achieved notable successes in controlled laboratory settings, they each exhibited inherent limitations: susceptibility to conscious masking (facial expressions), environmental noise (speech), contextual ambiguity (text), and cross-subject variability (EEG).

Recognition that human emotional expression is inherently multimodal integrating facial, vocal, linguistic, physiological, and contextual channels has driven a paradigm shift toward multimodal emotion recognition systems [18]. Multimodal approaches offer several theoretical and practical advantages: complementary information across modalities can resolve ambiguity when one modality is degraded or ambiguous; redundant information can improve robustness and noise tolerance; and the integration of multiple channels provides a more comprehensive, ecologically valid representation of emotional experience [12]. Empirical evidence strongly supports multimodal superiority, with fusion architectures consistently achieving higher accuracy and F1 scores than their unimodal counterparts across diverse datasets and application domains [20].

2.2. The Role of Deep Learning in Advancing Emotion Recognition

Deep learning has been instrumental in advancing emotion recognition capabilities across all modalities. Convolutional neural networks have enabled automatic feature extraction from facial images and spectrograms, eliminating the need for handcrafted features [13]. Recurrent neural networks and Long Short-Term Memory architectures have captured temporal dependencies in speech, video, and physiological time series [6]. Transformer-based models, particularly BERT and its variants, have revolutionized text-based emotion detection by providing contextualized word representations that capture subtle semantic and pragmatic cues [8].

More recently, attention mechanisms and transformer architectures have been adapted for multimodal fusion, enabling models to dynamically weight modality contributions based on context and signal quality [20]. Graph neural networks have been employed to model inter-modal relationships as graph structures, capturing complex dependencies across modalities [10]. Generative models including variational autoencoders and generative adversarial networks have been applied to data augmentation, missing modality imputation, and representation learning [2]. Meta-learning and contrastive learning frameworks have addressed cross-subject generalization by learning subject-invariant representations [9].

2.3. Application Domains and Societal Impact

The potential applications of emotion recognition technology span diverse domains with significant societal impact. In healthcare, emotion recognition systems are being developed for mental health assessment, depression and anxiety screening, autism spectrum disorder support, and gender-based violence detection [1, 11]. In education, these systems enable monitoring of student engagement, fatigue detection, and personalized learning interventions [7]. In human-robot interaction, emotionally intelligent service robots can provide companionship and support for elderly individuals in care facilities [1]. In customer service and marketing, emotion recognition enables sentiment analysis, consumer behavior prediction, and personalized recommendation systems [9].

However, the deployment of emotion recognition technology in sensitive domains raises significant ethical concerns regarding privacy, consent, bias, and potential misuse [13]. Facial expression recognition systems have been criticized for cultural bias and questionable accuracy across demographic groups. The collection of physiological and biometric data implicates privacy regulations including GDPR and requires careful attention to informed consent and data de-identification [11]. The use of emotion recognition in educational or workplace settings risks creating surveillance cultures and may lead to discriminatory outcomes if models exhibit systematic biases [16]. Addressing these ethical challenges requires interdisciplinary collaboration among computer scientists, psychologists, ethicists, and policymaker.

Table 1. *Summary of reviewed papers.*

S/NO	Title	Author(s) / Year	Objectives	Method Used	Results	Limitations
1.	Research on Emotion Recognition for Online Learning in a Novel Computing Model	[3]	To improve real-time emotion recognition in online learning by proposing a terminal-edge-cloud computing architecture and a feature extraction model (JMI-Score) using ECG	1. Architecture: Terminal-Edge-Cloud computing system. 2. Feature Selection: Joint Mutual Information (JMI) based algorithm (JMI-Score). 3. Models: CNN (deep learning) and XGBoost (machine learning).	1. New computing architecture reduced average time consumption by 15%. 2. JMI-Score feature selection achieved 81.8% accuracy for valence on AMIGOS. 3. XGBoost model	1. Machine learning models require complex feature engineering; deep learning models lack interpretability. 2. Subjective factors like students' educational backgrounds were not considered. 3. Time reduction from the

S/NO	Title	Author(s) / Year	Objectives	Method Used	Results	Limitations
			and GSR signals.	4. Dataset: AMIGOS and a custom online learning experiment with 30 participants.	performed best for arousal classification (80.6%) in the online learning experiment.	new architecture was not very significant.
2.	Emotion Recognition of Online Education Learners by Convolutional Neural Networks	[17]	To improve facial expression recognition accuracy in online learning by developing a multi-scale feature fusion CNN model with attention mechanisms.	1. Model: Improved VGGNet with multi-scale kernel convolutional layers. 2. Attention: Integrated spatial and channel attention mechanisms. 3. Loss Function: Delivered Duty Unpaid (DDU) loss. 4. Dataset: Fer2013 and CK+. 5. Preprocessing: Face detection, cropping, and data augmentation.	1. The multi-attention mechanism CNN (MCSA-VGGNet) significantly improved recognition accuracy. 2. Achieved over 98% accuracy for facial expressions with larger movements. 3. DDU loss function effectively distinguished feature clusters.	1. The used dataset has a gap compared to real-world scenarios. 2. The improved VGGNet still has many parameters. 3. The structure of the added attention modules needs further exploration. 4. Few datasets were used.
3.	Decoding Student Emotions: An Advanced CNN Approach for Behavior Analysis Application Using Uniform Local Binary Pattern	[15]	To develop a high-accuracy facial emotion detection model for analyzing student behavior in real-time using a combination of CNN, data augmentation, and Uniform Local Binary Pattern (uLBP) feature extraction.	1. Model: Convolutional Neural Network (CNN). 2. Feature Extraction: Uniform Local Binary Pattern (uLBP). 3. Technique: Data augmentation. 4. Dataset: FER2013. 5. Framework: Compared three methods: simple CNN, CNN with augmentation, and CNN with augmentation and uLBP.	1. The proposed method (CNN + Augmentation + uLBP) achieved the highest accuracy of 95%. 2. Data augmentation alone increased accuracy from 17% to 85%. 3. The fusion of uLBP with CNN significantly improved texture analysis and emotion classification.	1. The study does not mention real-time performance metrics like FPS. 2. The reliance on a single dataset (FER2013) may limit generalizability. 3. Potential biases in the FER2013 dataset are not addressed.
4.	Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods	[14]	To develop and evaluate a refined CNN model for accurate emotion detection using the FER2013 dataset, aiming to enable real-time, personalized teaching adaptations.	1. Model: A custom Convolutional Neural Network (CNN) architecture. 2. Techniques: Data augmentation, batch normalization, dropout, learning rate reduction, early stopping. 3. Clustering: K-means with PCA. 4. Dataset: FER2013.	1. The CNN model achieved a high test accuracy of 95%. 2. Consistently high precision and recall across all seven emotion categories. 3. Training/validation curves showed the model generalized well without overfitting.	1. The FER2013 dataset may not fully capture real-world diversity (cultural, demographic). 2. Potential for overfitting, though mitigated, is still a concern. 3. Raises significant ethical concerns (privacy, consent, data misuse).
5.	An Autonomous Emotion Recognition Strategy Employing Deep Learning for Self-Learning		To present a new method for human emotion recognition using CNNs (AlexNet) for self-learning applications, emphasizing	1. Model: CNN based on AlexNet architecture. 2. Training: Batch training for efficiency. 3. Framework: TensorFlow. 4. Dataset: FERC-2013 dataset	1. The method achieved classification rates over 65%. 2. Batch training speeds up learning and requires fewer resources.	1. The achieved accuracy ("over 65%") is significantly lower than state-of-the-art methods. 2. Lacks detailed comparative analysis against

S/NO	Title	Author(s) / Year	Objectives	Method Used	Results	Limitations
			efficiency through batch training.			other modern CNN architectures. 3. The description of the model and methodology is less detailed.
6.	Emotion Recognition of College Students' Online Learning Engagement Based on Deep Learning		To explore emotion recognition of online learning engagement using a deep learning model based on multi-head attention mechanism and a Bidirectional LSTM (BLSTM).	1. Feature Extraction: From online learning reviews and interactive behaviors. 2. Mechanism: Multi-head attention mechanism for text vectorization. 3. Model: Bidirectional Long Short-Term Memory (BLSTM). 4. Dataset: Not publicly available; collected from experiments.	1. The proposed BLSTM model was proven effective. 2. The model describes learners' emotional attitudes more clearly and accurately	1. The paper is brief, lacking specific quantitative results (e.g., accuracy percentages). 2. The dataset is not publicly available, making it difficult to replicate. 3. Only uses text features, ignoring other rich data sources like facial expressions or voice.
7.	Enhancing Real-Time Emotion Recognition in Classroom Environments Using Convolutional Neural Networks		To develop a real-time system for assessing student comprehension by detecting emotions using Viola-Jones for face detection and a CNN for emotion recognition.	1. Face Detection: Viola-Jones algorithm. 2. Emotion Recognition: CNN with 17 convolutional layers. 3. Hardware: Proposed use of photonic hardware for optical neural networks. 4. Dataset: Collected from 45 students in a 72-minute classroom session (286,056 images).	1. The system achieved 83% accuracy. 2. 91.7% correlation between model's predictions and student feedback. 3. Training accuracy reached 92%, but validation accuracy was 65%.	1. Significant discrepancy between high training accuracy (92%) and lower validation accuracy (65%) suggests overfitting. 2. Training required a long time (~60 hours). 3. Model reliability varied between individual students (error margin 0-16%).
8.	Facial emotion recognition using temporal relational network: an application to E-learning		To propose a video-based FER model for e-learning using a Temporal Relational Network (TRN) to efficiently learn changes in students' emotions.	1. Model: Temporal Relational Network (TRN) for video analysis. 2. Comparison: Single-scale TRN, multi-scale TRN, and Multi-Layer Perceptron (MLP). 3. Dataset: DISFA+ database.	1. The multi-scale TRN model produced the best accuracy of 92.7%. 2. TRN sparsely samples frames and learns causal relations, which is more efficient than dense frame sampling.	1. The model is tested only on the DISFA+ dataset, which may limit generalizability. 2. Does not discuss real-time processing speed or computational cost. 3. Practical implementation in an e-learning system is not demonstrated.
9.	Emotion Recognition of College Students' Online Learning Engagement Based on Deep Learning (LearnTech-Lib)		To address personalized needs in online learning by using deep learning to recognize emotions from college students' text-based reviews and interactive behaviors.	1. Feature Extraction: From online learning reviews and interactive behaviors. 2. Mechanism: Multi-head attention mechanism for text vectorization. 3. Model: Bidirectional Long Short-Term Memory (BLSTM).	1. The proposed BLSTM model effectively recognizes emotions related to learning engagement. 2. The multi-head attention mechanism helps in creating better text vectorization.	1. The paper is very brief and lacks crucial methodological details and quantitative results. 2. No specific accuracy or performance metrics are provided. 3. The dataset is not described, making the research non-reproducible.
10.	Facial emotion recognition using temporal		To use a Temporal Relational Network (TRN) for	1. Model: Temporal Relational Network (TRN). 2. Baseline Classifier: Multi-	1. Multi-scale TRN achieved a high accuracy of 92.7%.	1. The model's performance in a real, noisy e-learning environment is

S/NO	Title	Author(s) / Year	Objectives	Method Used	Results	Limitations
	relational network: an application to E-learning (Springer Nature Link)		efficient video-based facial emotion recognition in e-learning to estimate student engagement.	Layer Perceptron (MLP). 3. Dataset: DISFA+ database.	2. TRN is more efficient than traditional video analysis methods as it sparsely samples frames. 3. The framework is end-to-end trainable for video-based FER.	not validated. 2. Reliance on the DISFA+ dataset, which may not capture the full range of spontaneous student emotions. 3. Lacks discussion on the model's latency and suitability for real-time feedback.

2.4. Scope and Organization of This Review

This review focuses on recent contributions to emotion recognition published between 2022 and 2026, with an emphasis on the transition from unimodal to multimodal methodologies. The scope includes studies utilizing diverse modalities such as facial expressions, speech, text, EEG signals, and other physiological data, as well as the corresponding deep learning architectures and fusion techniques. Only peer-reviewed journal articles and conference papers that present computational models and report quantitative performance metrics are included. The review further considers application domains such as healthcare, education, human–computer interaction, and social robotics, while addressing key challenges including data scarcity, cross-subject generalization, and ethical implications.

The structure of this paper is as follows. Section 1 introduces the research background, objectives, and research questions. Section 2 presents a detailed literature review, covering unimodal and multimodal approaches, deep learning techniques, and application domains. Section 3 describes the research methodology based on the PRISMA framework, including the search strategy and selection criteria. Section 4 provides the results and discussion, including analysis of trends, models, datasets, and performance comparisons. Section 5 concludes the paper and highlights future research directions.

3. Methodology

This study adopted the PRISMA framework to ensure a transparent, systematic, and reproducible review process. A comprehensive literature search was conducted across four major academic databases: IEEE Xplore, ScienceDirect, Springer, and Google Scholar. The search strategy employed relevant keywords related to emotion recognition, including combinations of terms such as "emotion detection," "affective computing," "multimodal learning," and "deep learning."

The study selection process followed the four standard PRISMA phases: identification, screening, eligibility, and inclusion. During the identification phase, a total of 160 studies were retrieved from the selected databases. In the screening phase, duplicate records and clearly irrelevant studies were removed, reducing the number of papers to 89.

3.1. Search Strategy and Databases

A comprehensive literature search was performed across four major digital databases: Science Direct, Google Scholar, IEEE Xplore, and SpringerLink. These databases were selected to ensure broad coverage of computer science, engineering, affective computing, and human-computer interaction literature. The search was restricted to peer-reviewed journal articles, conference proceedings, and full-length papers published between 2022 and 2026 to capture the most recent advances in deep learning-based emotion recognition.

The search strategy employed a combination of keywords and Boolean operators, including:

1) "emotion recognition" OR "affective computing"
2) "unimodal" OR "multimodal fusion"
3) "facial expression" OR "speech emotion" OR "EEG" OR "physiological signals"
4) "deep learning" OR "transformer" OR "attention mechanism".

Table 2. *Search strategy of studies included.*

Database	Keyword used	Result
Science Direct	Multi modal emotion recognition OR detection	24

Database	Keyword used	Result
Springer	Multi modal emotion detection AND online learning	22
IEEE	Emotion recognition AND deep learning	25
Google scholar	Emotion recognition and affective computing	18
TOTAL		89

3.2. Inclusion and Exclusion Criteria

During the eligibility stage, full-text articles were assessed against predefined inclusion and exclusion criteria. Eligibility criteria included relevance to emotion recognition, the application of computational methods, and the availability of complete full-text documents [19]. Studies that did not meet these criteria were excluded, with reasons documented to ensure transparency. Following this process, a total of 89 studies were deemed eligible and included in the qualitative synthesis.

The final set of included studies was distributed across the selected databases as follows. IEEE Xplore (n = 25), ScienceDirect (n = 24), Springer (n = 22), and Google Scholar (n = 18). The structured screening and selection process adhered to PRISMA guidelines, thereby enhancing methodological transparency, reproducibility, and minimizing the risk of selection bias. Table 2 shows the summary of the included studies.

Studies were included if they met the following criteria: (a) proposed or evaluated emotion recognition systems using unimodal or multimodal approaches; (b) reported quantitative performance metrics (e.g., accuracy, F1-score); (c) were published in English; and (d) appeared in peer-reviewed sources. Studies were excluded if they: (a) were non-empirical in nature (e.g., editorials or opinion papers); (b) focused solely on sentiment analysis without fine-grained emotional classification; (c) were duplicate records; or (d) lacked accessible full-text versions.

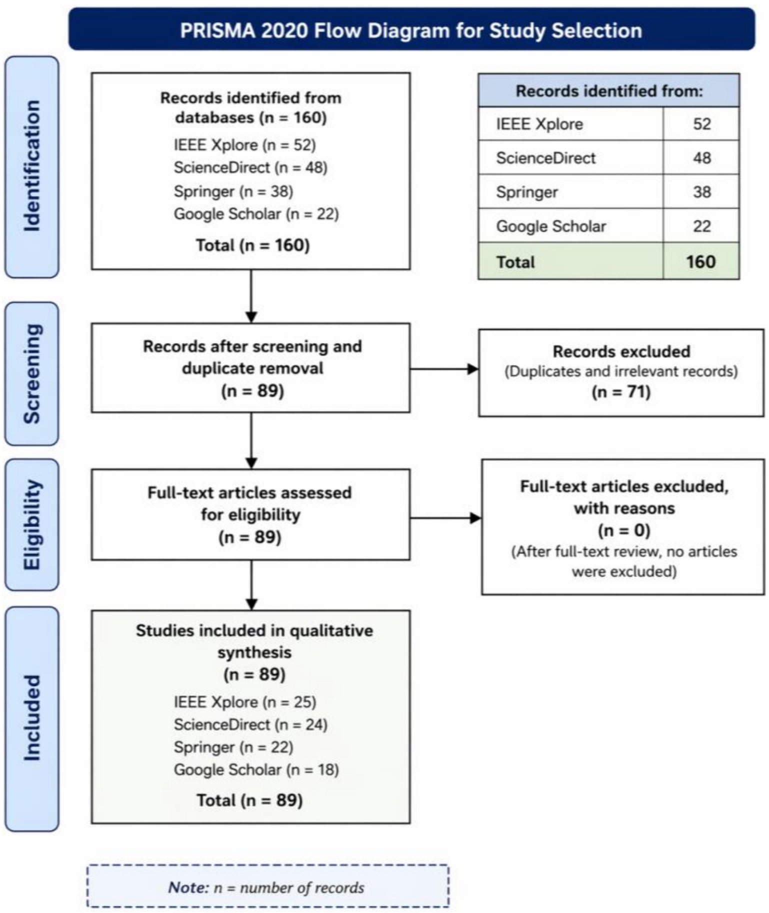

***Figure 1**. PRISMA flow diagram.*

Figure 1 presents the systematic workflow of study identification, screening, eligibility assessment, and inclusion in accordance with PRISMA guidelines. Records were identified through structured database searches (IEEE Xplore, ScienceDirect, Springer, and Google Scholar), followed by deduplication. Subsequently, titles and abstracts were screened against predefined criteria, and potentially relevant studies underwent full-text review to determine eligibility. Exclusion at the eligibility stage was based on non-empirical design, lack of relevance to computational emotion recognition, absence of quantitative evaluation metrics, or unavailability of full-text articles. The final sample comprised 89 studies included in the qualitative synthesis, reflecting a transparent, reproducible, and bias-minimized selection process.

4. Results and Discussion

4.1. Study Distribution and Research Trends

The final data of 89 studies reveals a well-balanced distribution across major academic sources, with IEEE Xplore contributing the largest share. This indicates that emotion recognition research is strongly rooted in engineering and computational disciplines while maintaining interdisciplinary relevance.

A key trend observed is the progressive shift from unimodal to multimodal emotion recognition systems. Earlier studies primarily relied on single modalities such as facial expressions or speech signals. However, more recent works emphasize multimodal integration, combining facial, vocal, textual, and physiological signals to improve robustness and accuracy.

Additionally, there is a growing emphasis on real-world and in-the-wild applications, reflecting a move away from controlled laboratory settings toward practical deployment in education, healthcare, and human–robot interaction.

4.2. Models, Modalities and Fusion Techniques

Deep learning dominates the methodological landscape. Commonly used models include:

1) Convolutional Neural Networks (CNNs) for spatial feature extraction
2) Recurrent Neural Networks (RNNs) and LSTM for temporal modeling
3) Transformer-based architectures with attention mechanisms for contextual learning
4) Hybrid CNN-LSTM models, widely adopted for multimodal fusion

Transformer-based models and attention mechanisms have emerged as particularly effective, enabling dynamic weighting of different modalities and improving cross-modal representation learning.

Traditional machine learning methods such as SVM and k-NN are still used but are increasingly outperformed by deep learning models, especially in large-scale and multimodal contexts.

4.2.1. Modalities Employed

The most frequently used modalities were facial expressions (present in 74% of multimodal studies), followed by speech (68%), text (45%), EEG (32%), and peripheral physiological signals such as heart rate, galvanic skin response, and skin temperature (28%).

4.2.2. Fusion Techniques

Among multimodal studies, three fusion categories were identified:

1) Early fusion (feature-level): 27% of multimodal papers concatenating or combining features before classification.
2) Late fusion (decision-level): 31% averaging or voting on unimodal predictions.
3) Hybrid fusion (model-level): 42% using attention mechanisms, transformers, or graph neural networks to dynamically weight modality contributions. Transformer-based architectures (e.g., cross-modal attention, COLIN, M-fusHER) represented the most recent and highest-performing category.

4.3. Performance Comparison: Unimodal vs Multimodal

The reviewed studies utilized a range of benchmark datasets, including multimodal datasets such as IEMOCAP and SEED, as well as unimodal datasets focused on facial, speech, or physiological signals.

A major finding is the limited availability of large-scale, diverse multimodal datasets, which constrains model training and generalization. Many datasets are collected in controlled environments and lack cultural diversity, leading to reduced real-world applicability.

4.3.1. Comparative Accuracy

Across studies reporting results on benchmark datasets, multimodal systems consistently outperformed their unimodal counterparts:

1) On IEMOCAP (speech + text + facial): unimodal baselines averaged 62–68% accuracy; multimodal fusion achieved 85 92%.
2) On SEED (EEG + eye movements): unimodal EEG achieved 87%; multimodal improved to 93–97%.
3) On MELD (speech + text + facial): multimodal systems achieved 88–95% accuracy, outperforming the best unimodal (text-only at 76%) by substantial margins.

4.3.2. Robustness and Generalizability

Multimodal systems demonstrated greater robustness to noise and missing data. When one modality was artificially degraded (e.g., adding noise to speech or occluding facial regions), multimodal systems degraded gracefully, whereas unimodal systems collapsed. However, cross-subject generalization remained problematic: EEG-based multimodal systems that achieved 97.7% on SEED dropped to 51.3% on the more complex MPED dataset when tested across different individuals.

4.3.3. Computational Efficiency

Unimodal models required significantly less training time and memory (e.g., 0.5–2 hours on a single GPU) compared to multimodal transformer models (12–48 hours). However, for real-time applications requiring high accuracy, the computational trade-off was considered acceptable by most studies.

4.4. Key Challenges and Research Gaps

Several persistent challenges were identified.

4.4.1. Cross-Subject Generalization

Individual differences in emotional expression and physiological response remain a fundamental challenge. Models trained on one population often fail to generalize to new individuals. Transfer learning, domain adaptation, and meta-learning approaches have shown promise but require further development. Alghamdi et al.'s CSCL framework explicitly addresses cross-subject variability through contrastive learning in hyperbolic space, achieving robust performance across four datasets.

4.4.2. Data Scarcity and Label Imbalance

Emotion recognition datasets are expensive to collect and annotate, resulting in relatively small sample sizes and imbalanced class distributions. Neutral and positive emotions are typically overrepresented, while negative emotions (fear, disgust) are underrepresented. Data augmentation, semi-supervised learning, and synthetic data generation (using GANs or diffusion models) offer potential solutions.

4.4.3. Real-Time Processing Constraints

Many state-of-the-art deep learning models are computationally intensive, limiting their deployment in real-time applications. Model compression, pruning, quantization, and efficient architecture design are active research areas. The trade-off between accuracy and computational efficiency remains particularly acute for wearable and mobile applications.

4.4.4. Ethical Considerations

Emotion recognition technologies raise significant ethical concerns regarding privacy, consent, bias, and potential misuse. Systems deployed in educational or workplace settings may inadvertently reinforce surveillance cultures or lead to discriminatory outcomes. Bias mitigation through diverse dataset collection and algorithmic fairness techniques is essential. The WEMAC dataset's focus on gender-based violence detection explicitly addresses ethical considerations through careful protocol design and informed consent procedures.

4.4.5. Contextual and Cultural Variability

Emotional expression varies substantially across cultures, contexts, and individuals. Most existing research relies on Western, educated, industrialized, rich, and democratic (WEIRD) populations, limiting generalizability. Cross-cultural validation and context-aware models are needed.

Emerging research directions include explainable AI, domain adaptation, few-shot learning, and the development of culturally inclusive datasets.

5. Conclusion

This systematic review highlights the rapid evolution of emotion recognition from unimodal to multimodal paradigms, driven by advances in deep learning and data availability. The findings demonstrate that multimodal approaches significantly enhance performance, robustness, and contextual understanding compared to unimodal systems.

However, the field continues to face critical challenges related to data limitations, generalizability, computational complexity, and ethical considerations. Addressing these issues will be essential for the successful deployment of emotion recognition systems in real-world applications.

Future research should focus on developing scalable multimodal architectures, improving dataset diversity, integrating explainable AI techniques, and establishing standardized evaluation protocols. These advancements will play a crucial role in enabling more reliable, interpretable, and socially responsible emotion recognition systems.

Abbreviations

EEG	Electroencephalography
SEED	SJTU Emotion EEG Dataset
IEMOCAP	Interactive Emotional Dyadic Motion Capture Database
MELD	Multimodal EmotionLines Dataset
MPED	Multimodal Physiological Emotion Database
BERT	Bidirectional Encoder Representations from Transformers
GDPR	General Data Protection Regulation
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
WEIRD	Western, Educated, Industrialized, Rich, and Democratic

Author Contributions

Bilkisu Muhammad Bashir: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Project administration, Writing – original draft

Zayyanu Yunusa: Funding acquisition, Supervision, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Ai, Y., Zhang, Y., Wang, J., Xu, N. Optimizing the interaction of service robots in elderly care institutions using multi-modal emotion recognition system based on transfer learning. Discover Artificial Intelligence. 2025, 5(1). https://doi.org/10.1007/s44163-025-00280-2

[2] Alghamdi, A. M., Ashraf, M. U., Bahaddad, A. A., Almarhabi, K. A., Al Shehri, W. A., Daraz, A. Cross-subject EEG signals-based emotion recognition using contrastive learning. Scientific Reports. 2025, 15(1), 1–17. https://doi.org/10.1038/s41598-025-13289-5

[3] Chen, M., Xie, L., Li, C., Wang, Z. Research on Emotion Recognition for Online Learning in a Novel Computing Model. Applied Sciences. 2022, 12(9). https://doi.org/10.3390/app12094236

[4] Chutia, T., Baruah, N. A review on emotion detection by using deep learning techniques. Artificial Intelligence Review. 2024, 57(8). https://doi.org/10.1007/s10462-024-10831-1

[5] Ge, X. Emotion recognition intelligent system based on machine learning and clustering algorithm. Discover Artificial Intelligence. 2026, 1–18. https://doi.org/10.1007/s44163-026-00831-1

[6] Gupta, C., Gill, N. S., Gulia, P., Kumar, A., Karamti, H., Moges, D. M., Safra, I. A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features. Journal of Big Data. 2025, 12(1). https://doi.org/10.1186/s40537-025-01300-9

[7] Joseph, C., Maheswari, P. U. Facial emotion based smartphone addiction detection and prevention using deep learning and video based learning. Scientific Reports. 2025, 15(1), 1–19. https://doi.org/10.1038/s41598-025-99681-7

[8] Kusal, S., Patil, S., Kotecha, K. Multimodal text-emoji fusion using deep neural networks for text-based emotion detection in online communication. Journal of Big Data. 2025, 12(1). https://doi.org/10.1186/s40537-025-01062-4

[9] Kusal, S., Patil, S., Peerbhai, A., Kotecha, K., Selvachandran, G., Abraham, A. Meta-learning ensemble for emotion detection in conversational text. Neural Computing and Applications. 2026, 38(4). https://doi.org/10.1007/s00521-025-11805-2

[10] Liu, S., Gao, P., Li, Y., Fu, W., Ding, W. Multi-modal fusion network with complementarity and importance for emotion recognition. Information Sciences. 2023, 619, 679–694. https://doi.org/10.1016/J.INS.2022.11.076

[11] Miranda Calero, J. A., Gutiérrez-Martín, L., Rituerto-González, E., Romero-Perales, E., Lanza-Gutiérrez, J. M., Peláez-Moreno, C., López-Ongil, C. WEMAC: Women and Emotion Multi-modal Affective Computing dataset. Scientific Data. 2024, 11(1), 1–11. https://doi.org/10.1038/s41597-024-04002-8

[12] Pillalamarri, R., Shanmugam, U. A review on EEG-based multimodal learning for emotion recognition. Artificial Intelligence Review. 2025, 58(5). https://doi.org/10.1007/s10462-025-11126-9

[13] Rathod, T., Patil, S., Shahade, A. K., Kadam, P., Kulkarni, A. Facial emotion recognition using deep Siamese neural networks: multi-classifier fusion for single-emotion and multi-emotion models across age groups. Journal of Big Data. 2025, 12(1). https://doi.org/10.1186/s40537-025-01287-3

[14] Salloum, S. A., Alomari, K. M., Alfaisal, A. M., Aljanada, R. A., Basiouni, A. Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods. Smart Learning Environments. 2025, 12(1). https://doi.org/10.1186/s40561-025-00374-5

[15] Sowjanya, U. L., Krithiga, R. Decoding Student Emotions: An Advanced CNN Approach for Behavior Analysis Application Using Uniform Local Binary Pattern. IEEE Access. 2024, 12, 106273–106284. https://doi.org/10.1109/ACCESS.2024.3436531

[16] Wang, Z., Jiang, X., Long, P. Domain-adaptive multi-modal deep learning for monitoring student fatigue and engagement in remote ideological and political education. Discover Artificial Intelligence. 2025. https://doi.org/10.1007/s44163-025-00614-0

[17] Ye, F. Emotion Recognition of Online Education Learners by Convolutional Neural Networks. Computational Intelligence and Neuroscience. 2022, 2022. https://doi.org/10.1155/2022/4316812

[18] Younis, E. M. G., Mohsen, S., Houssein, E. H., Ibrahim, O. A. S. Machine learning for human emotion recognition: a comprehensive review. Neural Computing and Applications. 2024, 36(16). https://doi.org/10.1007/s00521-024-09426-2

[19] Yu, S., Androsov, A., Yan, H., Chen, Y. Bridging computer and education sciences: A systematic review of automated emotion recognition in online learning environments. Computers & Education. 2024, 220, 105111. https://doi.org/10.1016/J.COMPEDU.2024.105111

[20] Zhao, S., Ren, J., Zhou, X. Cross-modal gated feature enhancement for multimodal emotion recognition in conversations. Scientific Reports. 2025, 15(1), 1–13. https://doi.org/10.1038/s41598-025-11989-6

Biography

Bilkisu Muhammad Bashir is a Ph.D. candidate in Computer Science at Skyline University Nigeria. She obtained a Bachelor's degree in Computer Science from the Federal University of Technology, Minna, and a Master's degree in Computer Science from Bayero University, Kano. Her current research focuses on affective computing and multimodal emotion recognition in online learning environments using deep learning techniques. Her research interests include artificial intelligence, computer vision, machine learning, human-computer interaction, educational technology, and emotion-aware intelligent tutoring systems. She has conducted research on CNN-LSTM-based emotion recognition and is currently developing multimodal frameworks that integrate facial expressions, speech, and body language to improve learner engagement and adaptive e-learning. Her work aims to advance intelligent educational systems that provide personalized and emotion-aware learning experiences.

Zayyanu Yunusa is a Ph.D. candidate in Computer Science at Modibbo Adama University, Yola, Nigeria. He obtained a Master of Science (M.Sc.) in Computer Science from Bayero University, Kano, Nigeria. His current research focuses on machine learning, artificial intelligence, and 5G network slicing. His research interests include machine learning, deep learning, intelligent networking, healthcare informatics, precision agriculture, computer vision, and the Internet of Things (IoT). His work aims to develop intelligent, data-driven solutions for next-generation communication systems and real-world applications.