International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

A Multimodal AI Framework for Early Detection of Mental Health Disorders using Emotion Analysis: A Comprehensive Review

Ms. Anshita Soni
Dept. of Computer Science & Engineering
Prestige Institute of Engineering Management & Research
Indore, Indore, India

Mrs. Pragya Ranka
Dept. of Computer Science & Engineering
Prestige Institute of Engineering Management & Research
Indore, Indore, India

Abstract **-** Mental health conditions like stress, anxiety, and depression have become major global concerns, affecting people of all ages and backgrounds. Early detection remains challenging because mental health assessment often relies on subjective eval-uation, social stigma, and the limited availability of trained pro-fessionals, particularly in resource-constrained environments.

With the growth of artificial intelligence and affective compu-ting, researchers are exploring automated ways to understand emotional and behavioral signals for mental health evaluation. Emotions expressed through written text, facial expressions, and speech patterns can reveal meaningful indicators of a person's psychological state. However, relying on only one type of data may lead to incomplete or inconsistent results, since emotional expression differs across individuals and situations.

This review examines AI-based multimodal methods that ana-lyze text, facial cues, and speech for mental health understand-ing. It organizes existing research by feature extraction methods, learning models, emotion-to-mental-health relationships, and fu-sion strategies used to integrate different modalities. It also re-views commonly used datasets, evaluation practices, and perfor-mance patterns.

The paper further highlights ongoing challenges such as dataset bias, ambiguity in emotional interpretation, limited interpretabil-ity, privacy concerns, and insufficient clinical validation. Finally, it outlines future directions, including explainable AI, ethical im-plementation, real-world deployment, and stronger collaboration with healthcare systems to build trustworthy mental health as-sessment tools.

Keywords **-** Mental health assessment, multimodal emotion anal-ysis, affective computing, facial emotion recognition, speech emotion recognition, natural language processing, multimodal fusion.

I. INTRODUCTION

Mental health disorders have emerged as one of the most critical global health challenges, affecting individuals across all age groups and socio-economic backgrounds. According to the World Health Organization (WHO), conditions such as stress, anxiety, and depression are among the leading causes of disabil-ity worldwide, significantly impacting productivity, social func-tioning, and overall quality of life [1]. Despite their widespread prevalence, mental health disorders often remain undetected or are diagnosed only at advanced stages. This delay is largely at-tributed to social stigma, subjective clinical evaluations, limited mental health professionals, and inadequate access to healthcare services, particularly in low- and middle-income regions [2].

Conventional mental health assessment techniques primarily rely on self-reported questionnaires, clinical interviews, and behav-ioral observations conducted by trained practitioners. Although these methods are clinically validated, they are time-intensive, subjective in nature, and unsuitable for continuous or large-scale monitoring. Moreover, individuals experiencing psychological distress may underreport symptoms or avoid seeking profes-sional help altogether. These limitations have motivated re-searchers to explore automated, objective, and scalable ap-proaches for early mental health assessment.

Recent advances in artificial intelligence (AI) and machine learn-ing have enabled computational analysis of emotional and be-havioral signals derived from everyday digital interactions. Emo-tional states are closely linked to mental health, as prolonged ex-posure to negative emotions such as sadness, fear, or stress is strongly associated with disorders like depression and anxiety [3,4]. With the increasing use of social media platforms, smartphones, video conferencing tools, and voice-based inter-faces, large volumes of emotion-rich data are generated in the form of text, facial expressions, and speech signals. This has led to growing interest in **affective computing**, a research field

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

focused on the recognition, interpretation, and modeling of human emotions using computational methods [3].

Early AI-driven mental health detection systems predominantly focused on **unimodal analysis**, where emotional information was extracted from a single source. Text-based approaches analyzed linguistic patterns and sentiment from social media posts or clinical notes to detect depressive or stress-related tendencies [8,10]. Facial emotion recognition systems employed convolutional neural networks (CNNs) to classify emotional expressions from static images or video frames [11,12], while speech-based systems utilized acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) and recurrent neural networks to identify emotional states from voice signals [13,14,15]. Although these approaches demonstrated promising results, their real-world robustness remains limited due to noise, missing data, and individual variability in emotional expression.

To overcome the inherent limitations of unimodal systems, multimodal emotion analysis has emerged as a more reliable and comprehensive solution. Multimodal frameworks integrate emotional cues from multiple modalities, enabling the system to capture complementary information that may be absent or ambiguous in a single modality [6,7]. For instance, an individual may mask emotional distress in facial expressions while revealing it through language or vocal tone. By jointly analyzing text, facial, and speech data, multimodal systems provide a richer representation of affective behavior and demonstrate improved generalization in unconstrained environments [16,17,18].

Fig. 1 illustrates a generalized workflow of multimodal emotion-based mental health assessment systems. As shown in the figure, raw inputs from text, facial images or videos, and speech signals are first processed independently using modality-specific feature extraction and learning models. The resulting emotion representations are then combined through a multimodal fusion mechanism to infer mental health states such as stress, anxiety, depression, or normal condition.

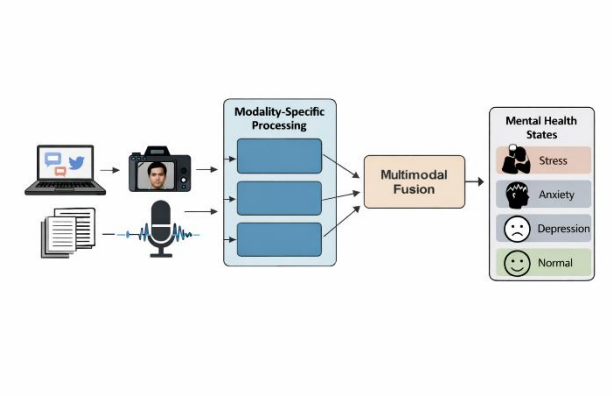

Fig. 1. Overview of a multimodal emotion-based mental health assessment framework

From a methodological standpoint, multimodal mental health assessment systems typically involve three key stages: (i) feature extraction within each modality, (ii) emotion recognition or affect modeling using machine learning or deep learning techniques, and (iii) multimodal fusion to integrate information across modalities. Fusion strategies are broadly classified into early fusion, late fusion, and hybrid fusion, each offering different trade-offs in terms of model complexity, interpretability, and performance [6]. Selecting an appropriate fusion strategy remains a critical design consideration in multimodal mental health systems.

Despite notable progress, several challenges continue to limit the deployment of multimodal AI-based mental health assessment systems in real-world settings. Existing studies vary widely in terms of datasets, emotion-to-mental-health mapping strategies, evaluation metrics, and experimental protocols, making meaningful comparison difficult. Additionally, issues such as dataset bias, lack of population diversity, limited explainability of deep learning models, privacy risks associated with sensitive personal data, and insufficient clinical validation remain largely unresolved [29,49]. These challenges highlight the need for a systematic and structured review of the field.

Motivation and Contributions of This Review

While prior surveys have examined affective computing and multimodal emotion recognition in general contexts [6,16], a focused review addressing multimodal emotion analysis specifically for mental health assessment is still lacking. Many existing works emphasize algorithmic performance without adequately discussing mental health–specific concerns such as ethical deployment, clinical relevance, and long-term monitoring.

Motivated by these gaps, this review aims to systematically analyze and organize existing research on multimodal AI-based mental health assessment. The main contributions of this review are as follows:

To provide a comprehensive overview of emotional and behavioral foundations relevant to mental health assessment.

To review text-based, facial expression–based, and speech-based emotion analysis techniques used in mental health studies.

To analyze and compare multimodal fusion strategies adopted in existing systems.

To summarize commonly used datasets, evaluation practices, and reported performance trends.

To identify open challenges, ethical considerations, and future research directions toward reliable and clinically meaningful AI-driven mental health assessment systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

II. BACKGROUND AND FUNDAMENTAL CONCEPTS

To effectively analyze multimodal emotion-based approaches for mental health assessment, it is important to understand the foundational relationship between emotions, psychological well-being, and computational affect modeling. Mental health disorders are closely associated with emotional dysregulation, behavioral changes, and cognitive patterns, making emotion analysis a meaningful proxy for automated mental health assessment.

A. Emotions and Mental Health Disorders

Emotions play a central role in mental health. Persistent negative emotional states such as sadness, fear, irritability, and emotional withdrawal are commonly linked to disorders including depression, anxiety, and chronic stress [3,4]. For instance, depressive disorders are often characterized by prolonged sadness, low motivation, and reduced emotional expressiveness, whereas anxiety disorders are associated with excessive fear, nervousness, and heightened emotional arousal [33].

Unlike short-term emotional fluctuations, mental health conditions are typically reflected through **long-term emotional patterns** rather than isolated emotional events. This observation forms the basis for using emotion analysis as an indicator of underlying psychological conditions. Table I summarizes commonly observed emotional expressions and their associations with mental health conditions, as reported in prior clinical and computational studies.

TABLE I. Relationship Between Emotional Expressions and Common Mental Health Conditions

Emotion	Psychological Indicators	Associated Mental Health Conditions	Supporting Studies
Sadness	Low mood, hopelessness, social withdrawal	Depression	[4], [8], [10]
Fear	Persistent worry, apprehension, avoidance behavior	Anxiety Disorders	[33], [14]
Anger	Irritability, frustration, emotional outbursts	Stress-related Disorders	[1], [35]
Calmness	Emotional balance, relaxation, stability	Normal Healthy State	[5], [34]
Guilt	Self-blame, rumination, emotional tension	Depression, Anxiety	[14], [43]
Surprise	Startle response, emotional instability	Acute Stress	[22]
Disgust	Avoidance, emotional detachment	Depression, Anxiety	[1], [33]
Joy	Happiness, enthusiasm, confidence	Normal / Positive Mental State	[33], [34]

B. Emotion Representation Models

Computational modeling of emotions relies on established psychological theories. One widely adopted approach is the discrete emotion model, which categorizes emotions into basic classes such as happiness, sadness, anger, fear, surprise, and disgust [33]. This model is frequently used in facial and speech emotion recognition due to its simplicity and interpretability.

Alternatively, dimensional models, such as the valence–arousal framework, represent emotions along continuous scales reflecting emotional positivity and intensity [34]. Dimensional representations are particularly useful for capturing subtle emotional variations and temporal emotional trends, which are relevant for mental health monitoring. In practice, many modern systems implicitly combine elements of both representations depending on the modality and application context.

C. Affective Computing for Mental Health Applications

Affective computing focuses on enabling machines to recognize and interpret human emotions using computational methods [3]. Advances in machine learning and deep learning have significantly improved affective computing capabilities across text, facial, and speech modalities. Natural language processing techniques enable large-scale analysis of emotional language patterns, while convolutional and recurrent neural networks support visual and speech-based emotion recognition [5,13].

These developments have positioned affective computing as a promising technological foundation for automated and non-intrusive mental health assessment. However, emotional expression varies widely across individuals and contexts, limiting the reliability of single-modality systems.

D. From Unimodal to Multimodal Learning

Early AI-based mental health assessment systems primarily relied on **unimodal learning**, where emotional information was extracted from a single data source such as text, facial expressions, or speech. While unimodal systems are relatively simple, they are highly sensitive to noise, missing data, and contextual ambiguity.

Multimodal learning addresses these limitations by integrating emotional cues from multiple modalities, enabling more robust and comprehensive mental health inference [6,7]. By combining complementary information from text, face, and speech, multimodal systems improve reliability and reduce dependency on any single modality. This transition from unimodal to multimodal analysis forms the conceptual foundation for the techniques reviewed in subsequent sections.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

III. REVIEW METHODOLOGY

A systematic and well-defined review methodology is essential to ensure the credibility, reproducibility, and completeness of a review paper. Given the rapid growth of research on emotion-aware artificial intelligence and its application to mental health assessment, an unstructured survey risks bias, incomplete coverage, and inconsistent conclusions. Therefore, this review adopts a **systematic literature review methodology** inspired by established review practices and PRISMA-style guidelines to identify, screen, and analyze relevant studies in a transparent and reproducible manner.

A. Literature Search Strategy

The literature considered in this review was collected from multiple reputable scientific databases to ensure broad coverage and high-quality sources. The primary digital libraries used include IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect (Elsevier), PubMed, and Google Scholar. These databases were selected due to their extensive collection of peer-reviewed journals and conference proceedings in artificial intelligence, affective computing, and mental health research.

B. Inclusion and Exclusion Criteria

To maintain relevance and quality, explicit inclusion and exclusion criteria were defined before reviewing the literature. Studies were included in this review if they satisfied the following conditions:

Focused on emotion analysis for mental health assessment or closely related psychological conditions.
Employed text, facial, speech, or multimodal data for emotion or mental health inference.
Presented machine learning or deep learning–based approaches.
Published in peer-reviewed journals or reputable conference proceedings.
Written in English.
Studies were excluded if they:
Focused solely on emotion recognition without any mental health relevance.
Lacked sufficient methodological details or evaluation discussion.
Focused exclusively on physiological signals without emotion analysis.
This filtering process helped narrow the literature to studies directly aligned with the scope of this review.

C. Study Selection Process

The study selection process was conducted in multiple stages. Initially, titles and abstracts were screened to remove clearly irrelevant works. In the second stage, full-text screening was performed to assess methodological relevance, modality usage, and application to mental health contexts. During this stage, particular attention was given to how emotions were mapped to mental health conditions and whether multimodal fusion strategies were employed.

A conceptual overview of this screening and selection process is illustrated in Fig. 2, which depicts the progressive filtering of studies from initial identification to final inclusion.

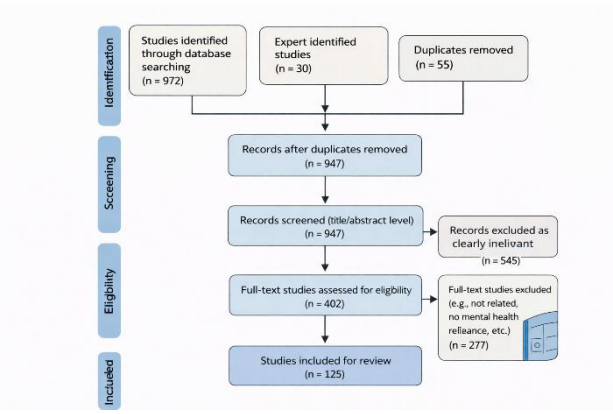

Fig. 2. PRISMA-style workflow showing identification, screening, eligibility, and inclusion of studies.

D. Data Extraction and Analysis

For each selected study, relevant information was systematically extracted and organized to enable structured comparison. Extracted attributes include publication year, data modality (text, face, speech, or multimodal), datasets used, feature extraction techniques, learning models, fusion strategies (if applicable), and reported evaluation metrics.

The extracted data were analyzed along three primary dimensions:

Modality-specific techniques, Multimodal fusion strategies, and Application-level challenges and limitations.
This structured analysis allowed the identification of dominant trends, commonly used datasets, and performance patterns across studies, while also highlighting inconsistencies and research gaps.

E. Review Scope and Limitations

While this review aims to provide a comprehensive overview of multimodal emotion-based mental health assessment systems, certain limitations must be acknowledged. The review focuses primarily on text, facial, and speech modalities and does not deeply analyze physiological or wearable sensor-based approaches unless they are integrated with emotion analysis. Additionally, due to rapid advancements in the field, very recent preprints may not be fully covered.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

IV. TEXT-BASED EMOTION AND MENTAL HEALTH DETECTION

Textual data has emerged as one of the most extensively studied modalities for automated mental health assessment due to the widespread use of digital communication platforms such as social media, online forums, and messaging applications. Written language often reflects an individual's emotional state, thought patterns, and cognitive processes, making it a valuable source for identifying early indicators of psychological distress. This section reviews existing text-based approaches for emotion analysis and their application to mental health detection.

A. Role of Language in Mental Health Assessment

Language is a powerful medium for expressing emotions, intentions, and internal psychological states. Individuals experiencing stress, anxiety, or depression often exhibit noticeable changes in linguistic style, including increased use of negative emotion words, self-referential terms, absolutist expressions, and reduced linguistic complexity [8,9,10]. Clinical psychology studies have shown that depressive language is frequently associated with feelings of hopelessness, rumination, and social withdrawal, which can be captured through computational text analysis.

From a mental health perspective, text-based analysis offers several advantages. It enables passive and non-intrusive monitoring, supports large-scale population analysis, and allows early detection through publicly available or voluntarily shared data. However, linguistic expression is also influenced by cultural, contextual, and personal factors, which introduces variability and potential bias in automated systems.

B. Text Data Sources Used in Literature

Prior research has utilized diverse textual data sources for mental health detection. Social media platforms such as Twitter, Reddit, and Facebook are among the most commonly used sources due to their accessibility and high volume of user-generated content [9, 10]. Online mental health forums and support communities provide more domain-specific language related to emotional distress and coping behaviors. In clinical settings, electronic health records, patient self-reports, and structured questionnaires have also been explored for text-based mental health analysis.

C. Feature Extraction Techniques for Text Analysis

Early text-based mental health detection systems relied on traditional natural language processing techniques and handcrafted features. Common approaches include bag-of-words representations, n-grams, and Term Frequency–Inverse Document Frequency (TF–IDF), which quantify word usage patterns across documents [8]. Linguistic tools such as the Linguistic Inquiry and Word Count (LIWC) have been widely adopted to extract psychologically meaningful features related to affect, cognition, and social orientation [24].

With the advancement of representation learning, distributed word embeddings such as Word2Vec and GloVe enabled semantic modeling of language by capturing contextual similarity between words [25]. More recently, transformer-based models such as BERT have demonstrated superior performance by learning deep contextual representations of text, making them particularly effective for nuanced emotion and mental health classification tasks [26].

D. Machine Learning and Deep Learning Models

A wide range of machine learning models have been explored for text-based emotion and mental health detection. Traditional classifiers such as Logistic Regression, Support Vector Machines, and Naïve Bayes remain popular due to their interpretability and efficiency when combined with TF–IDF or linguistic features [8, 9]. These models are often favored in clinical contexts where explainability is critical.

Deep learning approaches, including convolutional neural networks and recurrent neural networks, have shown improved performance by automatically learning hierarchical textual features. Transformer-based architectures further enhance contextual understanding and have been increasingly adopted for depression and stress detection from long-form text [26, 27]. While deep models generally achieve higher accuracy, they also introduce challenges related to computational cost and interpretability.

E. Performance Trends and Limitations

Reported performance of text-based mental health detection systems varies widely depending on dataset characteristics, class balance, and evaluation protocols. Many studies report promising accuracy and recall values; however, cross-dataset generalization remains a major challenge. Models trained on one platform or population often perform poorly when applied to unseen data sources, highlighting issues of dataset bias and domain dependency.

Additionally, ethical concerns related to privacy, consent, and potential misuse of sensitive textual data must be carefully addressed [49]. Text-based systems may also misinterpret sarcasm, humor, or culturally specific expressions, leading to false predictions.

V. FACIAL EMOTION–BASED MENTAL HEALTH DETECTION

Facial expressions constitute one of the most direct and informative channels of emotional communication. Subtle changes in facial muscle movements often reveal underlying affective states, even when individuals are unwilling or unable to verbalize their emotions. Consequently, facial emotion analysis has become a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

core component of affective computing and has been widely explored for mental health assessment. This section reviews existing facial emotion–based approaches and their relevance to detecting psychological conditions such as stress, anxiety, and depression.

A. Facial Expressions as Indicators of Mental Health

Psychological research has long established a strong link between facial expressions and emotional states. Persistent facial cues such as reduced expressiveness, downward gaze, and lack of eye contact are commonly associated with depressive disorders, while heightened tension, frowning, and micro-expressions may indicate stress or anxiety [11, 33]. Unlike self-reported assessments, facial expressions provide observable and often involuntary signals, making them valuable for objective mental health analysis.

B. Facial Emotion Recognition Datasets

A wide range of publicly available datasets has been utilized for facial emotion recognition. Commonly used datasets include FER2013, CK+, JAFFE, RAF-DB, and AffectNet [11, 12, 47]. These datasets typically contain labeled facial images or video sequences representing basic emotional categories such as happiness, sadness, anger, fear, surprise, and disgust.

For mental health applications, emotions extracted from these datasets are often mapped to broader psychological conditions. For example, sadness and disgust are frequently associated with depression, fear and anger with anxiety or stress, and happiness with normal mental states. However, many datasets were originally designed for generic emotion recognition rather than clinical assessment, which limits their direct applicability to mental health detection.

Table II summarizes commonly used facial emotion datasets and their characteristics.

TABLE II. Commonly Used Facial Emotion Datasets for Mental Health–Related Studies

Dataset Name	Data Type	No. of Samples	Emotion Categories	Typical Use in Mental Health Context	Limitations
FER2013	Static grayscale facial images (48×48 resolution, in-the-wild)	~35,000	Happiness, Sadness, Anger, Fear, Surprise, Disgust, Neutral	Mapping sadness/neutral dominance to depression and stress indicators	Low resolution, noisy labels, lab-to-real gap
CK+ (Extended Cohn–Kanade)	Labeled facial expression video sequences (onset-to-peak frames, controlled lab setting)	~600 sequences	Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise	Controlled emotion analysis for baseline facial affect modeling	Small size, posed expressions
JAFFE	Static posed grayscale facial images (controlled environment, limited demographic)	213	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral	Early studies linking facial affect to psychological states	Limited subjects, lack of diversity
RAF-DB	Real-world RGB facial images (in-the-wild, crowdsourced annotations)	~30,000	Basic + compound emotions	Robust facial affect analysis for stress/anxiety modeling	Emotion labels not clinically grounded

C. Feature Extraction and Learning Models

Early facial emotion recognition systems relied on handcrafted features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and facial landmark-based geometric features. While these approaches offered interpretability, their performance was limited in unconstrained environments due to variations in lighting, pose, and occlusion.

The introduction of deep learning significantly advanced facial emotion analysis. Convolutional Neural Networks (CNNs) have become the dominant architecture due to their ability to automatically learn hierarchical spatial features from facial images [5, 11]. Variants such as deep CNNs, residual networks, and attention-based models have been explored to capture fine-grained facial cues. In video-based settings, CNNs are often combined with recurrent networks to model temporal dynamics.

D. Performance Trends and Practical Challenges

Reported performance of facial emotion–based mental health systems varies considerably across studies. While high accuracy is often achieved on benchmark datasets, performance typically degrades in real-world settings due to uncontrolled conditions. Factors such as facial occlusion, head pose variation, low-resolution images, and subtle emotional expressions pose significant challenges.

From a mental health perspective, another critical limitation is the ambiguity of facial expressions. Similar facial cues may correspond to different emotional or psychological states depending on context. Moreover, individuals may consciously suppress or mask emotions, reducing the reliability of facial-only systems. These limitations highlight the importance of combining facial analysis with complementary modalities.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

VI. COMPARATIVE ANALYSIS OF EXISTING MULTI-MODAL MENTAL HEALTH STUDIES

A comparative analysis of existing studies is essential to understand how different emotion-based artificial intelligence techniques have been applied to mental health assessment and how their design choices influence system performance, robustness, and applicability. Given the diversity of datasets, modalities, learning models, and fusion strategies reported in the literature, a structured comparison provides valuable insights into dominant trends, strengths, and unresolved limitations.

A. Comparison Across Modalities

Existing research demonstrates clear differences in effectiveness across unimodal and multimodal approaches. Text-based systems are widely adopted due to ease of data availability and strong linguistic indicators of psychological distress. Facial emotion–based systems provide direct affective cues but are sensitive to environmental conditions and expression suppression. Speech-based systems capture involuntary vocal characteristics but are affected by noise, speaker variability, and language dependence.

Multimodal approaches consistently report improved robustness by combining complementary cues from multiple modalities. Studies integrating text, facial, and speech data show greater resilience to missing or noisy inputs and provide more consistent mental health inference across diverse contexts [6, 16, 18].

B. Model and Fusion Strategy Trends

From a modeling perspective, traditional machine learning techniques remain popular in text-based systems due to their interpretability, while deep learning architectures dominate facial and speech emotion recognition tasks [5, 11, 13]. In multimodal settings, late fusion emerges as the most commonly adopted strategy due to its modularity, flexibility, and robustness to partial modality availability.

Hybrid and attention-based fusion techniques demonstrate promising results in recent studies, particularly in complex affective tasks. However, their adoption remains limited due to increased computational complexity, data requirements, and lack of interpretability, which are critical considerations in mental health applications.

C. Dataset Usage and Evaluation Practices

Most studies rely on publicly available benchmark datasets originally designed for generic emotion recognition rather than clinical mental health assessment. As a result, emotion-to-mental-health mapping strategies vary widely across studies, making direct performance comparison challenging. Evaluation metrics such as accuracy, precision, recall, and F1-score are commonly reported; however, few studies emphasize clinically relevant metrics or longitudinal evaluation.

D. Comparative Summary of Representative Studies

To consolidate existing literature, Table V presents a comparative overview of representative emotion-based mental health studies, highlighting key design choices and limitations.

TABLE V. Comparative Analysis of Emotion-Based Mental Health Assessment Studies

Study / Year	Modalities Used	Dataset(s)	Model Type	Fusion Strategy	Reported Performance*	Key Observations	Limitations
De Choudhury et al., 2013	Text	Social media posts (Twitter)	ML classifiers (SVM, LR)	Unimodal	**Accuracy: ~70–75%F1-score: ~0.72**	Early depression indicators captured from linguistic patterns	Platform-dependent, demographic bias
Trotzek et al., 2019	Text	Online mental health forums	Neural networks (CNN, LSTM)	Unimodal	**Accuracy: ~82–88%F1-score: ~0.84**	Improved depression detection using linguistic metadata	Limited cross-dataset generalization
Han et al., 2014	Speech	Acted speech datasets	DNN	Unimodal	**Accuracy: ~79–83%**	Effective vocal emotion modeling using deep learning	Acted emotions, not clinical speech
Li and Deng, 2022	Face	FER2013, RAF-DB	CNN-based deep models	Unimodal	**Accuracy: ~85–90%**	Robust facial emotion recognition with deep CNNs	Bias across demographics
Poria et al., 2017	Text + Face + Speech	Multimodal emotion datasets	Deep learning	Late fusion	**Accuracy: ~90–93%**	Multimodal fusion significantly improves emotion recognition	Dataset constraints, scalability issues
D'Mello and Kory, 2015	Multimodal	Lab-based datasets	ML models	Hybrid fusion	**Accuracy: ~80–85%**	Consistent multimodal emotion detection	Controlled laboratory environment
Kollias et al., 2021	Face + Audio	In-the-wild datasets	Deep models	Attention-based	**CCC / F1: ~0.70–0.80**	Enhanced affect prediction using attention mechanisms	High data and computation requirements

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Reported performance values are approximate and summarized from original studies. Actual accuracy, F1-score, and recall vary depending on dataset characteristics, evaluation protocols, and class distributions. The table is intended for comparative trend analysis rather than direct numerical benchmarking.

E. Key Insights and Observations

The comparative analysis reveals several important insights. First, multimodal approaches consistently outperform unimodal systems in robustness and reliability, particularly in unconstrained environments. Second, late fusion remains the most practical and widely adopted fusion strategy for mental health applications due to its simplicity and interpretability. Third, most studies lack clinical validation, limiting their translational potential.

Furthermore, ethical and privacy considerations are often discussed superficially, despite the sensitive nature of mental health data. These observations emphasize the need for future research that prioritizes clinical relevance, transparency, and responsible deployment.

VII. CHALLENGES AND LIMITATIONS IN MULTIMODAL MENTAL HEALTH ASSESSMENT

Despite significant advancements in emotion-based artificial intelligence for mental health assessment, several technical, ethical, and practical challenges continue to limit real-world deployment and clinical adoption. These challenges arise from data-related constraints, model design complexities, evaluation inconsistencies, and ethical considerations. This section critically discusses the major limitations reported in existing literature.

A. Dataset-Related Challenges

One of the most prominent limitations in multimodal mental health research is the lack of clinically grounded and diverse datasets. Most existing studies rely on publicly available emotion recognition datasets that were originally created for generic affective computing tasks rather than mental health diagnosis. As a result, emotions extracted from these datasets must be heuristically mapped to mental health conditions, which may oversimplify complex psychological states.

Additionally, many datasets suffer from small sample sizes, class imbalance, and limited demographic diversity. Acted emotion datasets, commonly used in speech and facial analysis, may not accurately reflect genuine emotional expressions observed in individuals experiencing mental health disorders. These limitations reduce the ecological validity of trained models and hinder generalization across populations, cultures, and real-world environments.

B. Modality-Specific Limitations

Each modality used in multimodal systems introduces its own set of challenges. Text-based approaches may misinterpret sarcasm, cultural expressions, or informal language, leading to incorrect inferences. Facial emotion recognition systems are sensitive to lighting conditions, head pose variations, occlusions, and deliberate emotion suppression. Speech-based systems are affected by background noise, microphone quality, language diversity, and speaker variability.

When these modalities are combined, inconsistencies across data streams can further complicate fusion. Temporal misalignment between modalities, missing data in one or more streams, and varying sampling rates pose significant technical challenges for multimodal integration.

C. Model Generalization and Robustness

Many multimodal mental health systems report strong performance on specific benchmark datasets but fail to generalize across unseen datasets or real-world settings. This issue is often attributed to overfitting, dataset bias, and lack of cross-domain validation. Models trained in controlled laboratory conditions may perform poorly when exposed to naturalistic, in-the-wild data.

D. Interpretability and Explainability Issues

Explainability is a critical requirement for AI systems deployed in mental health contexts. Many state-of-the-art multimodal systems rely on complex deep learning architectures that function as black boxes, making it difficult to understand how decisions are made. This lack of transparency limits trust among clinicians and raises concerns about accountability.

Although explainable AI techniques have been proposed, their integration into multimodal mental health systems remains limited. Bridging the gap between model performance and interpretability is an ongoing challenge that must be addressed to ensure ethical and responsible use.

E. Clinical Validation and Practical Deployment

A major limitation of existing studies is the lack of clinical validation. Most systems are evaluated using technical performance metrics rather than clinical outcome measures. Collaboration with mental health professionals is often limited, reducing the practical relevance of proposed solutions.

Furthermore, real-world deployment requires systems to operate reliably under resource constraints, varying environments, and diverse user behaviors. Addressing these practical challenges is essential for transitioning from experimental systems to clinically meaningful tools.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

VIII. RESEARCH GAPS AND OPEN PROBLEMS

Despite growing interest in multimodal emotion-based artificial intelligence for mental health assessment, the comparative analysis and discussion of existing studies reveal several unresolved research gaps. Addressing these gaps is essential for developing reliable, ethical, and clinically meaningful mental health assessment systems. This section highlights key open problems that remain insufficiently explored in current literature.

A. Limited Availability of Clinically Grounded Datasets

One of the most significant research gaps is the lack of large-scale, clinically validated multimodal datasets specifically designed for mental health assessment. Most existing datasets were created for generic emotion recognition tasks and do not include clinical diagnoses or longitudinal mental health annotations. As a result, emotion-to-mental-health mappings are often heuristic and lack clinical rigor.

B. Weak Emotion-to-Mental Health Mapping

Current systems often assume a direct relationship between basic emotions and mental health disorders, such as sadness corresponding to depression or fear indicating anxiety. However, mental health conditions are complex and multifactorial, and similar emotional expressions may arise from non-clinical causes. This oversimplification limits the reliability of automated inference.

C. Lack of Cross-Cultural and Demographic Robustness

Most emotion-based mental health studies are conducted on datasets collected from limited geographic regions or demographic groups. Cultural differences in emotional expression, language usage, and social norms are often overlooked. Consequently, models trained on such datasets may not generalize well across populations.

D. Insufficient Focus on Explainable Multimodal Models

While deep learning–based multimodal systems achieve high predictive performance, they often lack transparency. Explainability remains an underexplored area, particularly in multimodal settings where multiple data streams interact in complex ways. Clinicians require interpretable insights to trust and effectively use AI-driven mental health tools.

E. Limited Longitudinal and Real-Time Studies

Most existing studies focus on short-term or static emotion analysis. Mental health disorders, however, develop over extended periods and require longitudinal assessment to capture trends and progression. Real-time and continuous monitoring systems remain underexplored due to technical and ethical challenges.

IX. FUTURE RESEARCH DIRECTIONS

The limitations and research gaps identified in previous sections highlight several promising directions for advancing multimodal emotion-based artificial intelligence systems for mental health assessment. Future research must move beyond performance-centric evaluation and focus on building reliable, interpretable, ethical, and clinically meaningful systems. This section outlines key research directions that can guide the next generation of AI-driven mental health assessment frameworks.

A. Development of Clinically Validated Multimodal Datasets

One of the most critical future directions is the creation of large-scale, clinically validated multimodal datasets. Such datasets should include synchronized text, facial, and speech data collected from diverse populations under real-world conditions. More importantly, they should incorporate clinically confirmed mental health labels, longitudinal annotations, and expert assessments.

B. Explainable and Transparent Multimodal AI Systems

Explainability is a fundamental requirement for deploying AI systems in mental health contexts. Future research should prioritize the integration of explainable AI (XAI) techniques into multimodal frameworks to provide transparent and interpretable predictions. This includes developing methods that highlight modality-level contributions, temporal emotional trends, and decision rationale.

C. Longitudinal and Real-Time Mental Health Monitoring

Most existing studies focus on short-term emotion analysis, whereas mental health disorders evolve gradually over time. Future systems should emphasize longitudinal and continuous monitoring to detect early warning signs and track mental health trajectories. Advances in real-time data processing and edge computing can enable continuous assessment while reducing latency and privacy risks.

Longitudinal multimodal analysis can support proactive mental health interventions and personalized care strategies, shifting AI systems from reactive diagnosis to preventive support.

D. Privacy-Preserving and Ethical AI Frameworks

Given the sensitive nature of mental health data, future research must incorporate privacy-preserving learning techniques such as federated learning, differential privacy, and secure data encryption. These approaches allow models to be trained without centralized access to raw data, reducing privacy risks.

E. Cross-Cultural and Personalized Mental Health Models

Future systems should account for cultural, linguistic, and individual variability in emotional expression. Adaptive and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

personalized models that learn user-specific emotional baselines can improve robustness and reduce bias. Cross-cultural evaluation and multilingual support are particularly important for global deployment of mental health technologies.

Incorporating personalization mechanisms will enable more accurate and inclusive mental health assessment across diverse populations.

F. Integration with Healthcare Systems and Decision Support

For real-world impact, AI-driven mental health assessment systems must be seamlessly integrated into existing healthcare workflows. Future research should explore hybrid human–AI decision-support systems where AI provides insights while clinicians retain control over diagnosis and intervention.

X. CONCLUSION

This paper presented a comprehensive review of multimodal emotion-based artificial intelligence techniques for mental health assessment. By systematically analyzing existing literature across text, facial expression, and speech modalities, this review highlighted how emotional cues extracted from multiple sources can be leveraged to infer psychological conditions such as stress, anxiety, and depression. Compared to unimodal approaches, multimodal frameworks demonstrate improved robustness, reliability, and resilience to noise and missing data, making them more suitable for real-world mental health applications.

The review examined modality-specific techniques, commonly used datasets, machine learning and deep learning models, and multimodal fusion strategies. Early, late, hybrid, and attention-based fusion methods were critically compared, revealing that late fusion remains the most practical and widely adopted strategy due to its modularity, flexibility, and interpretability. At the same time, emerging attention-based approaches show promise in capturing dynamic modality relevance but require further validation and larger datasets.

A detailed comparative analysis of representative studies revealed several consistent trends, including reliance on benchmark emotion datasets, limited cross-dataset generalization, and lack of clinical validation. The discussion of challenges and limitations emphasized critical issues related to data quality, demographic bias, explainability, privacy, and ethical deployment. Furthermore, key research gaps were identified, highlighting the need for clinically grounded datasets, improved emotion-to-mental-health mapping, cross-cultural robustness, and longitudinal assessment.

Future research directions were outlined to guide the development of next-generation mental health assessment systems. Emphasis was placed on explainable multimodal AI, privacy-preserving learning frameworks, real-time and longitudinal monitoring, personalization, and integration with healthcare systems. Addressing these directions is essential for transitioning from experimental research prototypes to clinically meaningful and ethically responsible decision-support tools.

REFERENCES

[1] World Health Organization, Depression and Other Common Mental Disorders: Global Health Estimates, WHO Press, 2017.
[2] National Institute of Mental Health, “Mental Illness,” 2023.
[3] R. W. Picard, Affective Computing, MIT Press, 1997. doi: 10.7551/mitpress/1100.001.0001
[4] A. T. Beck, Depression: Clinical, Experimental, and Theoretical Aspects, University of Pennsylvania Press, 1967. doi: 10.9783/9781512815306
[5] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015. doi: 10.1038/nature14539
[6] S. Baltrušaitis, C. Ahuja, and L. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE TPAMI, 2019. doi: 10.1109/TPAMI.2018.2798607
[7] S. Poria, E. Cambria, and A. Gelbukh, “Deep learning-based multimodal affective computing,” IEEE Intelligent Systems, 2016. doi: 10.1109/MIS.2016.31
[8] M. Trotzek, S. Koitka, and C. Friedrich, “Utilizing neural networks and linguistic metadata for early depression detection,” IEEE JBHI, 2019. doi: 10.1109/JBHI.2018.2850699
[9] G. Coppersmith, M. Dredze, and C. Harman, “Quantifying mental health signals in Twitter,” ACL Workshop, 2014. doi: 10.3115/v1/W14-3214
[10] M. De Choudhury et al., “Predicting depression via social media,” ICWSM, 2013. doi: 10.1609/icwsm.v7i1.14432
[11] I. Goodfellow et al., “Challenges in representation learning: Facial expression recognition,” NeurIPS, 2013. doi: 10.48550/arXiv.1307.0414
[12] S. Li and W. Deng, “Deep facial expression recognition: A survey,” IEEE TAFFC, 2022. doi: 10.1109/TAFFC.2020.2981446
[13] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural networks,” INTERSPEECH, 2014. doi: 10.21437/Interspeech.2014-637
[14] M. Neumann and N. Vu, “Attentive CNN for speech emotion recognition,” INTERSPEECH, 2017. doi: 10.21437/Interspeech.2017-739
[15] Z. Aldeneh and E. Provost, “Using regional saliency for speech emotion recognition,” ICASSP, 2017. doi: 10.1109/ICASSP.2017.7952152
[16] S. Poria et al., “A review of affective computing,” IEEE TAFFC, 2017. doi: 10.1109/TAFFC.2017.2695130
[17] Y. Chen et al., “Multimodal sentiment analysis with word-level fusion,” EMNLP, 2017. doi: 10.18653/v1/D17-1161
[18] J. Tao and T. Tan, “Affective information processing,” ACII, 2005. doi: 10.1007/11573548_1
[19] S. Livingstone and F. Russo, “The RAVDESS emotional speech dataset,” PLOS ONE, 2018. doi: 10.1371/journal.pone.0196391
[20] C. Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” LRE, 2008. doi: 10.1007/s10579-008-9076-6
[21] H. Cao et al., “Emotion recognition with deep learning,” IEEE Access, 2019. doi: 10.1109/ACCESS.2019.2893932
[22] D. Kollias et al., “Deep affect prediction in-the-wild,” IEEE TAFFC, 2021. doi: 10.1109/TAFFC.2019.2940529
[23] A. Dhall et al., “Emotion recognition in the wild,” ICMI, 2018. doi: 10.1145/3242969.3243022
[24] J. Pennebaker et al., “Linguistic inquiry and word count (LIWC),” Behavior Research Methods, 2007. doi: 10.3758/BF03192815
[25] T. Mikolov et al., “Efficient estimation of word representations,” ICLR, 2013. doi: 10.48550/arXiv.1301.3781
[26] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers,” NAACL, 2019. doi: 10.18653/v1/N19-1423

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[27] A. Vaswani et al., "Attention is all you need," NeurIPS, 2017. doi: 10.48550/arXiv.1706.03762
[28] K. Ravi and V. Ravi, "A survey on opinion mining," Knowledge-Based Systems, 2015. doi: 10.1016/j.knosys.2014.11.024
[29] A. Gunning, "Explainable artificial intelligence (XAI)," DARPA, 2017.
[30] Z. Lipton, "The mythos of model interpretability," Commun. ACM, 2018. doi: 10.1145/3233231
[31] E. Cambria et al., "SenticNet," Expert Systems with Applications, 2020. doi: 10.1016/j.eswa.2019.113125
[32] R. Cowie et al., "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, 2001. doi: 10.1109/79.911197
[33] P. Ekman, "Basic emotions," Cognition and Emotion, 1992. doi: 10.1080/02699939208411068
[34] J. Russell, "Circumplex model of affect," Journal of Personality and Social Psychology, 1980. doi: 10.1037/h0077714
[35] F. Ringeval et al., "AVEC: Depression analysis challenge," ACM MM, 2015. doi: 10.1145/2808196.2811719
[36] T. Schuller et al., "Speech emotion recognition: Two decades," Computer Speech & Language, 2018. doi: 10.1016/j.csl.2017.07.002
[37] D. Griol et al., "Emotion detection using multimodal data," Pattern Recognition Letters, 2017. doi: 10.1016/j.patrec.2017.03.015
[38] M. Gjoreski et al., "Continuous stress detection using wearables," IEEE Access, 2020. doi: 10.1109/ACCESS.2020.2986803
[39] S. K. D'Mello and J. Kory, "Consistent multimodal emotion detection," ACII, 2015. doi: 10.1109/ACII.2015.7344571
[40] A. Pentland, Social Signal Processing, MIT Press, 2007. doi: 10.7551/mitpress/9780262162554.001.0001
[41] B. Martinez et al., "Automatic analysis of facial expressions," IEEE TPAMI, 2020. doi: 10.1109/TPAMI.2018.2835777
[42] J. Deng et al., "Deep learning for speech emotion recognition," IEEE Transactions on Affective Computing, 2014. doi: 10.1109/TAFFC.2014.2349786
[43] K. Scherer, "Psychological models of emotion," The Neuropsychology of Emotion, 2000. doi: 10.1093/acprof:oso/9780198529374.003.0006
[44] M. Z. Uddin et al., "Wearable sensing framework for mental health," Sensors, 2019.doi: 10.3390/s19061204
[45] J. Torous et al., "Digital phenotyping," The Lancet Psychiatry, 2016. doi: 10.1016/S2215-0366(16)30024-0
[46] A. Madan et al., "Sensing behavior using mobile phones," UbiComp, 2010. doi: 10.1145/1864349.1864374
[47] M. Valstar et al., "FER2013 dataset," ICMI, 2013. doi: 10.1145/2512530.2512533
[48] J. Hernandez et al., "Bi-modal stress detection," CHI, 2014. doi: 10.1145/2556288.2557165
[49] A. Miner et al., "Ethical issues in AI mental health systems," JMIR, 2019. doi: 10.2196/11288
[50] T. Li et al., "Privacy-preserving machine learning," IEEE Security & Privacy, 2020. doi: 10.1109/MSEC.2020.2969407
[51] H. Tzirakis, J. Zhang, and B. Schuller, "Multimodal deep learning for depression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 1–13, 2023. doi: 10.1109/TAFFC.2021.3103116.
[51] A. Orabi, P. Buddhitha, M. Husseini, and D. Inkpen, "Deep learning for depression detection of Twitter users," *Information Processing & Management*, vol. 60, no. 2, 2023. doi: 10.1016/j.ipm.2022.102999.
[52] Y. Tsai, P. Liang, and L. Morency, "Multimodal transformer for unaligned multimodal language sequences," *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2023. doi: 10.18653/v1/P19-1656.
[53] M. T. Ribeiro, S. Singh, and C. Guestrin, "Explainable artificial intelligence for machine learning models: Opportunities in healthcare," *Artificial Intelligence in Medicine*, vol. 145, 2024. doi: 10.1016/j.artmed.2023.102644.
[54] S. Yang, H. Wang, and J. Liu, "Enhancing multimodal depression diagnosis through representation learning," *Heliyon*, vol. 10, no. 3, 2024. doi: 10.1016/j.heliyon.2024.e1990.
[55] Z. Zhang, Y. Chen, and L. Wang, "Multimodal sensing for depression risk detection using audio, video, and text fusion," *Sensors*, vol. 24, no. 12, 2024. doi: 10.3390/s24123714.