EuroVantage
Journal of Artificial Intelligence

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Multimodal AI for Early Detection of Depression and Anxiety Disorders

Dr. Vinodhini Ravikumar
Technical Founder at Mind Mosaic AI, Canada
Corresponding Email: vini@mindmosaicai.com

Abstract

Depression and anxiety disorders are among the most common mental health conditions globally, and if left untreated, they can have major emotional, social and economic impacts. The major methods of diagnosis are the clinical interview, self-report instruments, and observation evaluation; all of which can be subject to subjective bias and/or delayed reporting of symptoms. The recent developments of artificial intelligence (AI) have engendered novel possibilities for improved mental health screening by bringing together multimodal data sources. By combining information from textual content, speech patterns, facial expressions, physiological signals, and behavioral indicators, multimodal AI systems can capture a more comprehensive representation of an individual's psychological state.

This research focuses on the utilization of multimodal artificial intelligence (AI) methods in the detection of depression and anxiety disorders early. It summarizes the important data modalities, feature extraction techniques, fusion techniques, and deep learning architectures used in modern mental health assessment systems. The efficacy of multimodal models is also assessed in comparison to single-modality models, in particular in the improvement of predictive accuracy, robustness and diagnostic reliability. Further, some of the major problems with data privacy, model interpretability, dataset heterogeneity, and clinical deployment are covered. The results suggest significant potential for multimodal AI frameworks to assist clinicians in the objective, scalable and continuous monitoring of mental health. The authors' findings suggest that the adoption of these sophisticated multimodal learning methods, explainable AI systems, and personalized analytics can substantially impact the success of these EI strategies and potentially drive better mental healthcare outcomes.

Keywords: Multimodal Artificial Intelligence, Depression Detection, Anxiety Disorders, Mental Health Assessment, Deep Learning, Multimodal Fusion, Early Diagnosis, Clinical Decision Support Systems, Behavioral Analytics, Explainable AI

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

I. Introduction

Depression and anxiety disorders are some of the most common mental disorders in the world and impact people of all ages, socio-economic status and geographic locations. The psychological distress; lower quality of life; social dysfunction; lower productivity in the workplace; higher health utilization costs; are all associated with these disorders. A timely diagnosis and treatment can be crucial to better outcomes and to prevent symptoms from worsening into more serious psychiatric disorders. Traditional clinical diagnosis is mainly based on clinical interviews, self-reported questionnaires, and observational assessment, which may be subjective, time consuming, and reports biased. For this reason, objective and scalable technologies that can assist clinicians in early detection and monitoring of mental health disorders are becoming increasingly popular (Zafar et al., 2024; Pavlopoulos et al., 2024).

New technologies harnessing artificial intelligence (AI) have revolutionized mental health assessment by allowing for the automated analysis of complex behavioral, emotional, and physiological factors involved in depression and anxiety. Machine learning and deep learning algorithms show great promise in uncovering patterns in different types of data, like speech signals, facial expressions, text, physiology, or digital traces of behavior. AI-powered systems can provide opportunities for continuous monitoring and personalized mental health assessments, which complement more traditional clinical practices and improve access to mental health screening options (Barua et al., 2024; Kumar et al., 2024).

While these advances have occurred, numerous current AI diagnostic systems depend on just a single data modality, such as text, speech, or facial images. Unimodal solutions have yielded some success, but cannot adequately reflect the complexity of psychological states and emotions in humans. It is difficult to fully capture an individual's mental health status from any single data source because mental health disorders are expressed verbally, nonverbally, cognitively, and physiologically. As a result, researchers have increasingly turned toward multimodal AI frameworks that integrate information from multiple sources to improve diagnostic reliability and predictive accuracy (Mamidisetti & Reddy, 2022; Arioz et al., 2022).

Multimodal AI combines heterogeneous data streams such as audio recordings, facial videos, textual narratives, electroencephalogram (EEG) signals, wearable sensor measurements, and behavioral interactions into unified analytical frameworks. By leveraging complementary information across modalities, multimodal systems can identify subtle patterns that may be overlooked when individual modalities are analyzed independently. Previous studies have demonstrated that multimodal fusion techniques significantly enhance the detection of depression and anxiety by capturing both observable behaviors and underlying physiological responses. Deep learning architectures such as convolutional neural networks (CNNs), long short-term memory (LSTM) networks, transformer-

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

based models, and multimodal fusion frameworks have achieved improved classification performance compared with traditional approaches (Xie et al., 2022; Wang et al., 2021; Zhang et al., 2020).

The availability of specialized multimodal datasets has further accelerated research in this domain. The synchronized text, speech, video, and physiological signals found in datasets offer valuable resources for training and evaluating AI models for mental health assessment. For instance, the MMDA dataset has been created specifically to help identify depression and anxiety in the multimodal space, allowing for research into advanced fusion techniques and machine learning algorithms (Jiang et al., 2022). Alongside this, the multimodal-multisensor approach has raised the stakes for anxiety measurement with the use of physiological and contextual data from wearable and digital platforms to support more holistic assessments of emotional and behavioral states (Senaratne et al., 2022).

In recent years, novel multi-modal systems have emerged that incorporate large language model, facial expression recognition and reinforcement learning systems to enhance mental health screening and intervention. Several studies have shown the effectiveness of using a combination of textual and audio information for depression diagnosis and the newly developed approaches are investigating the incorporation of facial expression with LLM to improve predictive capabilities and interpretability (Mohammad & Al Mansoor, 2024; Sadeghi et al., 2024). Multimodal data-driven models for anxiety screening have also demonstrated the promise of intelligent technologies in facilitating proactive mental health care and personalized interventions (Mo et al., 2024; Pathirana et al., 2024).

In light of the swift advancements in multimodal AI technology and its increasing significance within mental health care, it is justifiable to investigate existing approaches, data modalities, fusion techniques, applications, and challenges. This study reviews and assesses the current status of multimodal AI for early recognition of depression and anxiety disorders, while outlining potential avenues for future research to create clinically viable, accurate, and explainable systems for assessing mental health. The current study seeks to build on existing research and technological progress to shed light on the potential for multimodal AI in facilitating timely, objective, and accessible mental health screening and intervention.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

II. Literature Review

2.1 Evolution of Artificial Intelligence in Mental Health Assessment

The growing burden of depression and anxiety disorders has intensified the search for effective methods capable of supporting early detection and intervention. Traditional diagnostic approaches rely heavily on clinical interviews, self-reported questionnaires, and behavioral observations, which may be influenced by subjective biases and delayed symptom reporting. Artificial intelligence (AI) has emerged as a promising solution by enabling automated analysis of behavioral, physiological, and emotional indicators associated with mental health conditions. Recent studies demonstrate that AI-based systems can identify subtle patterns that are often difficult for clinicians to detect through conventional assessment methods alone (Zafar et al., 2024; Pavlopoulos et al., 2024).

Initial AI applications in mental health focused primarily on single-modality data sources such as text, speech, facial expressions, or physiological signals. While these approaches achieved moderate success, researchers increasingly recognized that depression and anxiety manifest through multiple interconnected behavioral and biological dimensions. Consequently, multimodal AI systems have gained significant attention because they integrate information from diverse sources, providing a more comprehensive representation of an individual's mental state (Mamidisetti & Reddy, 2022; Arioz et al., 2022).

2.2 Multimodal Data Sources for Depression and Anxiety Detection

Multimodal AI systems utilize multiple forms of data to improve prediction accuracy and robustness. Common modalities include textual information, speech signals, facial expressions, physiological measurements, and behavioral indicators. Each modality contributes unique information regarding emotional and cognitive functioning.

Textual data are widely used for identifying linguistic markers associated with depression and anxiety. Features such as sentiment polarity, vocabulary usage, emotional expressions, and semantic patterns have demonstrated strong predictive value. Audio-based approaches analyze speech characteristics including pitch variation, speaking rate, vocal intensity, and pause frequency, which often reflect psychological distress (Mohammad & Al Mansoor, 2024).

Visual modalities primarily focus on facial expression analysis, eye movement tracking, and head posture recognition. Research has shown that facial micro-expressions and reduced emotional expressiveness can serve as indicators of depressive symptoms. Kumar et al. (2024) further demonstrated the effectiveness of combining facial expression recognition with electroencephalogram (EEG) signals for early depression identification. Physiological signals, including EEG recordings, heart rate variability, and wearable sensor measurements, provide

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

objective biomarkers capable of capturing neurophysiological changes associated with anxiety and depression (Senaratne et al., 2022).

Table 1. Major Modalities Used in Multimodal Depression and Anxiety Detection

Modality	Data Source	Key Features Extracted	Clinical Significance
Textual	Clinical interviews, questionnaires, social media posts	Sentiment, linguistic patterns, semantic features	Identifies emotional and cognitive disturbances
Audio	Speech recordings and conversations	Pitch, speech rate, pauses, vocal intensity	Detects vocal indicators of psychological distress
Visual	Facial videos and images	Facial expressions, eye gaze, head movements	Measures affective and behavioral changes
Physiological	EEG, heart rate sensors, wearables	Brain activity, heart rate variability, stress signals	Provides objective mental health biomarkers
Behavioral	Smartphone usage and activity logs	Interaction patterns, mobility, social engagement	Supports continuous mental health monitoring

2.3 Multimodal Fusion Techniques

A central component of multimodal AI systems is the fusion of heterogeneous data sources. Fusion techniques are generally categorized as early fusion, late fusion, and hybrid fusion approaches. Early fusion combines features extracted from different modalities before classification, enabling models to learn joint representations of multimodal information. Although effective, this approach often faces challenges associated with high-dimensional feature spaces and modality synchronization.

Late fusion independently processes each modality before combining individual predictions. This method provides greater flexibility when dealing with missing or incomplete data. Hybrid fusion techniques integrate the strengths of both approaches and have become increasingly popular in recent deep learning frameworks (Arioz et al., 2022).

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Among advanced multimodal architectures, convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, transformers, and attention-based mechanisms have demonstrated considerable effectiveness. Xie et al. (2022) proposed a CNN-LSTM fusion framework that successfully integrated multimodal information for depression and anxiety diagnosis, achieving improved classification performance compared with single-modality systems. Similarly, Zhang et al. (2020) developed a multimodal deep learning framework capable of combining visual, auditory, and textual cues for mental disorder recognition.

Recent advancements have further incorporated large language models (LLMs) into multimodal frameworks. Sadeghi et al. (2024) demonstrated that combining facial expression analysis with LLM-based textual understanding enhances depression detection performance by leveraging contextual and emotional information simultaneously.

2.4 Multimodal Datasets and Benchmark Resources

The development of reliable multimodal datasets has played a critical role in advancing AI-based mental health research. High-quality datasets enable researchers to train, validate, and benchmark predictive models under standardized conditions. One notable contribution is the Multimodal Depression and Anxiety (MMDA) dataset introduced by Jiang et al. (2022), which integrates textual, audio, and visual information specifically designed for depression and anxiety detection tasks.

The availability of multimodal datasets has facilitated comparative evaluation of different machine learning and deep learning architectures. However, several limitations remain, including limited sample sizes, demographic imbalance, cultural variability, and inconsistencies in data collection protocols. These challenges can affect model generalizability and hinder clinical deployment.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Table 2. Representative Multimodal Datasets and Associated Applications

Dataset/Study	Modalities Included	Primary Application
MMDA Dataset (Jiang et al.)	Text, Audio, Video	Depression and Anxiety Detection
Wang et al.	Face Video	Depression and Anxiety Diagnosis
Kumar et al.	Facial Expressions, EEG	Early Depression Detection
Sadeghi et al.	Facial Expressions, Textual Data	Depression Detection Using LLMs
Mo et al.	Multimodal Sensor and Behavioral Data	Anxiety Screening

2.5 Applications of Multimodal AI in Depression and Anxiety Detection

Several studies have demonstrated the effectiveness of multimodal AI systems in improving early diagnosis. Wang et al. (2021) showed that face-video-based multimodal fusion approaches could effectively distinguish individuals experiencing depression and anxiety symptoms. Mo et al. (2024) introduced a multimodal data-driven framework that integrated multiple sources of information to enhance anxiety screening accuracy.

Similarly, Mohammad and Al Mansoor (2024) developed a unified multimodal deep learning model that combined textual and speech-based features for depression diagnosis. Their findings indicated that multimodal integration significantly improved classification performance compared to unimodal approaches. Pathirana et al. (2024) further expanded the application of multimodal AI through reinforcement learning-based emotion recognition systems designed to support mental health promotion and intervention.

Barua et al. (2024) highlighted the growing importance of AI-assisted tools for detecting depression and anxiety among adolescents, emphasizing their potential role in preventing suicidal ideation through early intervention. These findings collectively suggest that multimodal AI can serve as an effective decision-support tool for clinicians and mental health practitioners.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

2.6 Research Gaps and Future Research Needs

Despite substantial progress, several challenges continue to limit the widespread adoption of multimodal AI systems in mental healthcare. One major issue involves data privacy and ethical concerns arising from the collection of sensitive behavioral and physiological information. Additionally, many existing models suffer from limited interpretability, making it difficult for clinicians to understand and trust algorithmic predictions (Zafar et al., 2024).

Another challenge involves dataset heterogeneity and demographic bias. Many datasets are collected from relatively small and homogeneous populations, reducing the generalizability of trained models across diverse clinical settings. Furthermore, multimodal systems often require substantial computational resources and complex data synchronization processes, which may hinder real-world implementation (Senaratne et al., 2022).

Current research increasingly emphasizes explainable AI, privacy-preserving learning frameworks, large language model integration, and continuous monitoring through wearable technologies. Addressing these challenges is expected to improve the reliability, scalability, and clinical applicability of multimodal AI systems for early detection of depression and anxiety disorders.

III. Methodology

3.1 Research Design

This study adopted a systematic analytical research design to investigate the effectiveness of multimodal artificial intelligence (AI) techniques for the early detection of depression and anxiety disorders. The methodology focused on evaluating how multiple data modalities, including textual, audio, visual, physiological, and behavioral signals, can be integrated to improve the identification of mental health conditions. The research design was selected because depression and anxiety are complex disorders that manifest through diverse cognitive, emotional, and behavioral patterns that cannot be adequately captured using a single data source (Mamidisetti & Reddy, 2022; Senaratne et al., 2022).

A comparative framework was developed to analyze existing multimodal AI architectures, feature extraction approaches, and fusion mechanisms used in depression and anxiety detection systems. The study emphasized the examination of multimodal deep learning models, large language model-assisted frameworks, and hybrid machine learning approaches that have demonstrated promising performance in mental health assessment (Sadeghi et al., 2024; Mohammad & Al Mansoor, 2024).

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

3.2 Data Sources and Modalities

The study examined multimodal datasets and data acquisition approaches commonly employed in depression and anxiety detection research. Data modalities were categorized into textual, audio, visual, physiological, and behavioral sources.

Textual data included clinical interview transcripts, social media content, patient narratives, and questionnaire responses. Audio data consisted of speech recordings from interviews and conversational interactions, from which acoustic features such as pitch, energy, speech rate, and pauses were extracted. Visual data comprised facial videos and images used to capture emotional expressions, gaze behavior, and facial movement patterns. Physiological signals included electroencephalography (EEG), heart rate variability, and other sensor-derived indicators of emotional and cognitive states. Behavioral data incorporated smartphone interactions, activity patterns, and multimodal emotion recognition outputs (Jiang et al., 2022; Kumar et al., 2024; Pathirana et al., 2024).

The MMDA dataset and other benchmark multimodal datasets were considered representative sources due to their inclusion of synchronized audio, visual, and textual information for depression and anxiety assessment (Jiang et al., 2022). Previous studies have demonstrated that combining these modalities provides richer contextual information and improves predictive performance compared to unimodal systems (Arioz et al., 2022; Wang et al., 2021).

3.3 Data Preprocessing

To ensure consistency and reliability, each modality underwent modality-specific preprocessing procedures prior to feature extraction.

For textual data, preprocessing involved tokenization, stop-word removal, text normalization, and semantic embedding generation. Natural language processing techniques were applied to identify linguistic markers associated with depression and anxiety, including sentiment polarity, emotional expressions, and lexical complexity (Mohammad & Al Mansoor, 2024).

Audio signals were processed through noise reduction, segmentation, and normalization procedures. Acoustic features such as Mel-frequency cepstral coefficients (MFCCs), pitch variation, speech intensity, and pause duration were extracted to capture vocal indicators of psychological distress (Xie et al., 2022).

Visual data preprocessing involved face detection, alignment, normalization, and frame extraction. Facial landmarks and expression-related features were identified using computer vision techniques to detect emotional states associated with depression and anxiety (Kumar et al., 2024).

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Physiological signals were filtered to remove artifacts and noise before extracting relevant temporal and frequency-domain characteristics. Behavioral datasets were cleaned and standardized to facilitate integration with other modalities (Senaratne et al., 2022).

3.4 Feature Extraction and Representation

Feature extraction was performed independently for each modality to obtain meaningful representations of emotional, cognitive, and behavioral indicators.

Textual features were generated using transformer-based language representations and contextual embeddings capable of capturing semantic relationships within clinical and conversational data. Audio features were extracted using speech analysis techniques designed to identify variations in vocal patterns associated with depressive and anxious symptoms. Visual features were derived from facial expression recognition models that quantified emotional responses through facial action units and movement patterns (Zhang et al., 2020).

Physiological features included EEG spectral characteristics, heart rate variability measures, and other biosignal-derived indicators of stress and emotional regulation. Behavioral features incorporated activity patterns, engagement metrics, and emotion recognition outputs derived from digital interactions (Mo et al., 2024; Pathirana et al., 2024).

The extracted features were transformed into standardized vector representations to facilitate multimodal fusion and model training.

3.5 Multimodal Fusion Framework

The proposed analytical framework employed multimodal fusion to integrate information from diverse data sources. Three major fusion strategies were examined: early fusion, late fusion, and hybrid fusion.

Early fusion combined features from all modalities into a unified representation before classification. This approach enabled the model to learn interactions among heterogeneous features during training. Late fusion processed each modality independently and combined prediction outputs at the decision level. Hybrid fusion integrated both feature-level and decision-level information to leverage the strengths of each strategy (Arioz et al., 2022).

Deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models were investigated due to their ability to capture complex temporal and contextual relationships across modalities. CNN-LSTM frameworks were particularly emphasized because of their demonstrated effectiveness in multimodal depression and anxiety diagnosis (Xie et al., 2022). Recent multimodal large language model frameworks

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

incorporating facial expression analysis were also considered due to their enhanced contextual reasoning capabilities (Sadeghi et al., 2024).

3.6 Model Development

The multimodal classification framework consisted of four sequential stages: data acquisition, feature extraction, multimodal fusion, and classification.

Following feature integration, machine learning and deep learning models were trained to classify individuals into depression, anxiety, or non-clinical categories. Model architectures included CNNs for spatial feature learning, LSTMs for temporal sequence modeling, and transformer-based models for contextual representation learning (Zhang et al., 2020; Mohammad & Al Mansoor, 2024).

The framework was designed to support scalable mental health screening and decision support applications. Model optimization techniques such as regularization, dropout, and hyperparameter tuning were incorporated to reduce overfitting and improve generalization performance (Mo et al., 2024).

3.7 Performance Evaluation

Model performance was evaluated using widely accepted classification metrics employed in mental health prediction research. Accuracy was used to measure overall classification performance, while precision and recall assessed the correctness and completeness of predictions. The F1-score provided a balanced measure of classification effectiveness, particularly for imbalanced datasets.

Sensitivity and specificity were also examined because of their clinical significance in identifying individuals with depression and anxiety while minimizing false-positive diagnoses. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was utilized to evaluate the discriminative ability of the multimodal models across different classification thresholds (Barua et al., 2024; Pavlopoulos et al., 2024).

Comparative analyses were conducted between unimodal and multimodal approaches to determine the contribution of each modality and fusion strategy. Performance outcomes reported in previous studies consistently indicated that multimodal systems outperform single-modality models in detecting depression and anxiety due to their ability to capture complementary behavioral and emotional information (Wang et al., 2021; Xie et al., 2022).

3.8 Ethical Considerations

Given the sensitive nature of mental health data, ethical considerations were incorporated into the research framework. Particular attention was given to privacy protection, informed consent, data

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

anonymization, and responsible AI deployment. The study also considered concerns related to algorithmic bias, fairness, and explainability, which remain critical challenges in AI-assisted mental health assessment (Zafar et al., 2024; Barua et al., 2024).

Ensuring transparency and interpretability of multimodal AI models is essential for fostering trust among clinicians, patients, and healthcare organizations. Therefore, explainable AI principles were considered an important component of future multimodal mental health screening systems.

IV. Results and Discussion

4.1 Performance of Multimodal AI Models for Depression and Anxiety Detection

The reviewed studies consistently demonstrate that multimodal artificial intelligence systems outperform traditional single-modality approaches in the early detection of depression and anxiety disorders. By integrating heterogeneous data sources such as speech, facial expressions, text, physiological signals, and behavioral indicators, multimodal frameworks capture complementary manifestations of mental health conditions that are often missed when relying on a single data stream. This capability is particularly important because symptoms of depression and anxiety are expressed through complex interactions between verbal communication, emotional expression, cognitive patterns, and physiological responses (Mamidisetti & Reddy, 2022).

Several studies have reported substantial improvements in classification accuracy through multimodal fusion. Zhang et al. (2020) proposed a multimodal deep learning framework that combined visual, acoustic, and textual information to recognize mental disorders more effectively than unimodal methods. Similarly, Xie et al. (2022) demonstrated that integrating multiple modalities within a CNN-LSTM architecture improved the diagnosis of depression and anxiety by capturing both spatial and temporal behavioral patterns. Wang et al. (2021) further showed that facial video-based multimodal fusion can identify emotional irregularities associated with depressive and anxious states, highlighting the value of visual information in mental health assessment.

Recent developments have extended multimodal systems to include advanced deep learning architectures and large language models. Mohammad and Al Mansoor (2024) introduced a unified multimodal deep learning framework that combines textual and speech features for depression diagnosis, demonstrating robust predictive capabilities. Likewise, Sadeghi et al. (2024) integrated facial expression analysis with large language models, illustrating how generative AI can enhance contextual understanding of emotional and linguistic cues. These findings collectively indicate that multimodal learning provides richer representations of psychological states and improves early screening performance.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Table 3. Comparison of Representative Multimodal AI Models for Depression and Anxiety Detection

Study	Modalities Used	AI Technique	Key Findings
Zhang et al. (2020)	Text, Audio, Visual	Multimodal Deep Learning	Improved recognition of mental disorders through integrated feature learning
Wang et al. (2021)	Face Video	Multimodal Fusion Framework	Enhanced identification of depression and anxiety symptoms
Xie et al. (2022)	Audio and Visual Data	CNN-LSTM	Improved diagnostic performance through temporal-spatial feature extraction
Mohammad & Al Mansoor (2024)	Text and Speech	Unified Deep Learning Model	Robust depression diagnosis using multimodal fusion
Mo et al. (2024)	Multiple Behavioral Signals	Data-Driven Framework	Effective anxiety screening through multimodal analysis
Sadeghi et al. (2024)	Facial Expressions and Text	Large Language Models + Vision Models	Enhanced depression detection through contextual multimodal reasoning

The findings summarized in Table 3 reveal a clear trend toward the adoption of deep neural architectures capable of learning complex relationships among heterogeneous data sources. Across studies, multimodal models consistently achieved superior detection capabilities compared with systems based solely on text, audio, or visual information.

4.2 Benefits of Multimodal AI in Early Mental Health Screening

A major advantage of multimodal AI is its ability to improve diagnostic robustness by leveraging complementary information from diverse sources. Individuals suffering from depression or anxiety may exhibit subtle symptoms that are difficult to identify through a single modality. For example,

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

speech abnormalities may coexist with altered facial expressions and changes in language usage. Integrating these indicators allows AI systems to generate a more comprehensive assessment of mental health status (Senaratne et al., 2022).

Another significant benefit is improved resilience to incomplete or noisy data. If one modality becomes unavailable or produces low-quality information, other modalities can compensate, thereby maintaining predictive performance. Arioz et al. (2022) reported that multimodal systems generally exhibit greater stability and reliability than unimodal approaches. Similarly, Mo et al. (2024) demonstrated that combining multiple behavioral and emotional indicators improves anxiety screening effectiveness across diverse populations.

Multimodal AI also supports continuous and non-invasive monitoring. Advances in wearable devices, smartphone sensors, and conversational AI platforms have enabled the collection of real-time behavioral and physiological data. Such capabilities facilitate early intervention by detecting emerging symptoms before they develop into severe mental health conditions. Pathirana et al. (2024) further demonstrated that multimodal emotion recognition systems can support personalized mental health promotion through adaptive learning mechanisms.

Table 4. Key Benefits of Multimodal AI for Depression and Anxiety Detection

Benefit	Description	Supporting Studies
Improved Accuracy	Combines complementary information from multiple sources	Xie et al. (2022); Zhang et al. (2020)
Robustness	Maintains performance despite missing or noisy data	Arioz et al. (2022)
Early Detection	Identifies subtle behavioral and emotional changes before clinical manifestation	Mo et al. (2024); Kumar et al. (2024)
Continuous Monitoring	Enables real-time mental health assessment through sensors and digital platforms	Pathirana et al. (2024)
Personalized Assessment	Supports individualized prediction and intervention strategies	Senaratne et al. (2022)

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Clinical Decision Support	Assists healthcare professionals in screening and diagnosis	Pavlopoulos et al. (2024)

The evidence suggests that multimodal AI systems can significantly enhance mental health screening by improving both predictive performance and clinical utility. Their ability to process diverse information sources positions them as valuable tools for supporting healthcare professionals in identifying at-risk individuals.

4.3 Challenges and Limitations of Existing Multimodal Systems

Despite encouraging results, several challenges continue to hinder the widespread adoption of multimodal AI in mental healthcare. One of the most significant concerns involves data privacy and security. Mental health information often contains highly sensitive personal data, including speech recordings, facial images, physiological measurements, and behavioral patterns. The collection, storage, and processing of such information raise ethical concerns regarding confidentiality and informed consent (Zafar et al., 2024).

Another challenge relates to dataset quality and representativeness. Existing multimodal datasets often contain limited demographic diversity, resulting in potential algorithmic bias and reduced generalizability. Jiang et al. (2022) highlighted the importance of comprehensive datasets such as MMDA for improving depression and anxiety detection; however, larger and more diverse datasets remain necessary to ensure fairness and applicability across populations.

Model interpretability also remains a critical issue. Although deep learning architectures provide strong predictive performance, many operate as black-box systems whose decision-making processes are difficult to explain. Healthcare practitioners may be reluctant to rely on diagnostic recommendations that cannot be clearly justified. Barua et al. (2024) emphasized that explainability is essential for fostering trust and facilitating clinical adoption of AI-assisted mental health tools.

Furthermore, most existing studies are conducted in controlled research environments, which may not accurately reflect real-world clinical conditions. Variations in recording quality, environmental noise, patient behavior, and cultural differences can affect system performance during deployment. As a result, additional clinical validation is required before multimodal AI systems can be routinely integrated into healthcare workflows.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Table 5. Challenges and Potential Solutions in Multimodal Mental Health

Challenge	Impact on Detection Systems	Potential Solution
Privacy and Security Concerns	Reduced user trust and regulatory barriers	Federated learning, secure data sharing frameworks
Dataset Bias and Limited Diversity	Reduced generalizability and fairness	Development of larger and more representative datasets
Lack of Explainability	Limited clinical acceptance	Explainable AI and interpretable deep learning models
Data Heterogeneity	Difficult integration of multiple modalities	Standardized multimodal data processing frameworks
Real-World Deployment Constraints	Performance degradation in practical settings	Extensive clinical validation and field testing
Computational Complexity	Increased resource requirements	Efficient model optimization and edge AI solutions

The challenges identified in Table 5 indicate that technical performance alone is insufficient for successful implementation. Future developments must address ethical, operational, and clinical considerations to ensure responsible adoption of multimodal AI technologies.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

4.4 Discussion

The collective evidence indicates that multimodal AI has emerged as one of the most promising approaches for early detection of depression and anxiety disorders. By integrating diverse forms of behavioral, physiological, visual, and linguistic information, these systems provide a more comprehensive representation of mental health status than traditional assessment methods. The progression from conventional machine learning approaches toward deep learning, multimodal fusion networks, and large language model-enhanced architectures has substantially improved predictive capabilities (Sadeghi et al., 2024; Mohammad & Al Mansoor, 2024).

The findings also reveal that multimodal systems are increasingly moving beyond simple diagnostic classification toward continuous monitoring and personalized mental healthcare. Emerging frameworks leverage wearable devices, emotion recognition technologies, and adaptive learning mechanisms to support proactive mental health management (Pathirana et al., 2024; Pavlopoulos et al., 2024). These developments align with the growing emphasis on preventive healthcare and early intervention strategies.

Nevertheless, significant challenges remain regarding privacy, interpretability, fairness, and clinical validation. Addressing these issues will be essential for translating promising research outcomes into trustworthy and widely adopted healthcare solutions. Future studies should focus on explainable multimodal architectures, larger cross-cultural datasets, and real-world deployment evaluations to maximize the clinical impact of AI-assisted mental health screening systems.

V. Future Directions

The rapid advancement of multimodal artificial intelligence has created new opportunities for enhancing the early detection of depression and anxiety disorders. Although existing systems have demonstrated promising performance by integrating facial expressions, speech characteristics, textual information, physiological signals, and behavioral indicators, several research challenges remain. Future developments are expected to focus on improving model robustness, scalability, interpretability, and clinical applicability.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

5.1 Integration of Large Language Models and Advanced Multimodal Architectures

Recent developments in large language models (LLMs) present significant opportunities for improving multimodal mental health assessment. Traditional multimodal frameworks primarily rely on deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and CNN-LSTM hybrid models for feature extraction and classification (Xie et al., 2022; Zhang et al., 2020). Future systems may incorporate LLMs capable of understanding contextual linguistic information while simultaneously processing visual and acoustic cues. Such architectures can facilitate richer representation learning, improved symptom interpretation, and enhanced conversational screening capabilities. Emerging studies combining facial expressions with language-based reasoning demonstrate the potential of these models to improve depression detection accuracy and provide more comprehensive assessments of emotional states (Sadeghi et al., 2024).

5.2 Expansion of Wearable and Multisensor Monitoring Systems

Another important direction involves the integration of wearable technologies and multisensor platforms for continuous mental health monitoring. Current multimodal approaches often rely on controlled datasets and clinical assessments, limiting their applicability in real-world environments (Senaratne et al., 2022). Future research should explore the incorporation of smartwatches, fitness trackers, electroencephalogram (EEG) devices, and smartphone sensors to collect physiological and behavioral information in a nonintrusive manner. The combination of heart rate variability, sleep patterns, physical activity levels, facial expressions, and speech characteristics could enable continuous risk assessment and early intervention. Such developments would support proactive mental healthcare by identifying subtle behavioral changes before clinical symptoms become severe (Kumar et al., 2024; Pavlopoulos et al., 2024).

5.3 Personalized and Adaptive Mental Health Assessment

Mental health disorders manifest differently across individuals due to variations in age, gender, culture, personality traits, and environmental factors. Consequently, future multimodal AI systems should move beyond generalized predictive models toward personalized and adaptive frameworks. Reinforcement learning and adaptive learning mechanisms may enable systems to continuously refine predictions based on individual behavioral patterns and emotional responses (Pathirana et al., 2024). Personalized models could provide more accurate assessments by accounting for unique psychological baselines and longitudinal behavioral changes. Such individualized approaches may significantly improve screening accuracy while reducing false-positive and false-negative outcomes.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

5.4 Personalized and Adaptive Mental Health Assessment

Mental health disorders manifest differently across individuals due to variations in age, gender, culture, personality traits, and environmental factors. Consequently, future multimodal AI systems should move beyond generalized predictive models toward personalized and adaptive frameworks. Reinforcement learning and adaptive learning mechanisms may enable systems to continuously refine predictions based on individual behavioral patterns and emotional responses (Pathirana et al., 2024). Personalized models could provide more accurate assessments by accounting for unique psychological baselines and longitudinal behavioral changes. Such individualized approaches may significantly improve screening accuracy while reducing false-positive and false-negative outcomes.

5.5 Development of Larger and More Diverse Multimodal Datasets

The availability of high-quality datasets remains a critical factor influencing model performance and generalizability. Existing datasets often contain limited demographic diversity, relatively small sample sizes, or constrained recording conditions, which can affect the reliability of AI models when deployed across broader populations (Jiang et al., 2022; Arioz et al., 2022). Future efforts should prioritize the creation of large-scale, geographically diverse, and clinically validated multimodal datasets that include textual, visual, audio, physiological, and behavioral information. Standardized benchmarking datasets would facilitate more consistent model evaluation and promote reproducibility across studies. Increased dataset diversity would also help reduce algorithmic bias and improve model fairness across different demographic groups.

5.6 Explainable and Ethical Artificial Intelligence

Despite significant progress in prediction accuracy, the limited interpretability of deep learning models remains a major barrier to clinical adoption. Healthcare professionals often require transparent explanations to understand why a system has generated a particular prediction. Future research should emphasize explainable AI techniques capable of identifying which multimodal features contribute most significantly to depression and anxiety detection outcomes (Zafar et al., 2024). Furthermore, ethical considerations related to privacy, informed consent, data security, and algorithmic fairness must remain central to future system development. Privacy-preserving machine learning approaches and secure data-sharing mechanisms may help address concerns associated with the collection and analysis of sensitive mental health information (Barua et al., 2024).

5.7 Clinical Translation and Real-World Deployment

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

While many multimodal AI systems have achieved encouraging results in experimental settings, relatively few have been integrated into routine clinical practice. Future investigations should focus on validating these technologies through large-scale clinical trials and real-world implementation studies. Collaboration among clinicians, psychologists, data scientists, and healthcare organizations will be essential for developing clinically reliable and operationally feasible systems. Integration with telemedicine platforms, electronic health records, and digital mental health services may further enhance accessibility and facilitate earlier intervention for individuals at risk of depression and anxiety disorders (Mo et al., 2024; Wang et al., 2021).

Overall, future research is expected to move toward intelligent, explainable, personalized, and continuously adaptive multimodal AI systems capable of supporting comprehensive mental health assessment. Advances in multimodal fusion, large language models, wearable sensing technologies, and ethical AI frameworks will play a pivotal role in transforming depression and anxiety screening from episodic clinical evaluations to proactive and data-driven mental healthcare ecosystems.

VI. Conclusion

The increasing prevalence of depression and anxiety disorders has created a need for effective, scalable and timely screening tools that can aid clinical decision making and early intervention. This study explored the potential of the multimodal Artificial Intelligence in early detection of depression and anxiety disorders, and the use of a multimodal representation of the mental health of the individual, using multiple modalities of data, such as textual content, speech signals, facial expressions, physiological signals, and behavioral signals, can offer an accurate picture of the mental health status of an individual which is more complex than can conventional methods. The results show that the multimodal AI systems consistently outperform unimodal approaches by learning complementary patterns from various sources of information and, consequently, achieve the best diagnostic performance, robustness and reliability (Zhang et al., 2020; Xie et al., 2022).

The review also identified that the AI systems' ability to detect subtle signs of depression and anxiety has significantly improved with the development of deep learning architectures, multimodal fusion strategies, and data-driven screening frameworks. Some models employing a tri-modal fusion of audio, visual, and text data and physiological data have shown good results in detecting early signs and behavioral indicators of mental illness (Mo et al., 2024; Mohammad & Al Mansoor 2024). The development and evaluation of increasingly sophisticated diagnostic models have also been supported by available specialized multimodal datasets and benchmark resources (Jiang et al., 2022). Moreover, the studies that have recently included large language models, facial expression analysis, and multimodal emotion recognition provide new possibilities for developing adaptive and context-aware mental health assessment systems (Sadeghi et al., 2024; Pathirana et al., 2024).

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

Although these developments have occurred, there are still a few issues that hinder the use of these technologies in the clinics. Data privacy, ethical governance, transparency of models, demographic bias and generalizability are still important concerns that need careful attention. Existing studies emphasize the importance of explainable AI, secure data-sharing mechanisms, and rigorous clinical validation to ensure that multimodal diagnostic systems can be trusted and effectively integrated into healthcare environments (Arioz et al., 2022; Senaratne et al., 2022). In addition, the need for larger, more diverse datasets and standardized evaluation protocols remains critical for improving model robustness across different populations and healthcare settings (Mamidisetti & Reddy, 2022; Barua et al., 2024).

Overall, multimodal AI represents a transformative approach to mental health assessment, offering substantial potential for the early identification of depression and anxiety disorders. By leveraging complementary information from multiple data sources, these systems can support clinicians in delivering more objective, personalized, and proactive care. Continued progress in multimodal learning, explainable artificial intelligence, and human-centered system design is expected to further strengthen the effectiveness of AI-assisted mental health screening and contribute to improved patient outcomes and mental healthcare accessibility (Zafar et al., 2024; Pavlopoulos et al., 2024; Kumar et al., 2024; Wang et al., 2021)

References

1. Jiang, Y., Zhang, Z., & Sun, X. (2022, August). MMDA: a multimodal dataset for depression and anxiety detection. In *International Conference on Pattern Recognition* (pp. 691-702). Cham: Springer Nature Switzerland.
2. Xie, W., Wang, C., Lin, Z., Luo, X., Chen, W., Xu, M., ... & Cheng, M. (2022). Multimodal fusion diagnosis of depression and anxiety based on CNN-LSTM model. *Computerized Medical Imaging and Graphics*, *102*, 102128.
3. Mo, H., Hui, S. C., Liao, X., Li, Y., Zhang, W., & Ding, S. (2024). A multimodal data-driven framework for anxiety screening. *IEEE transactions on instrumentation and measurement*, *73*, 1-13.
4. Arioz, U., Smrke, U., Plohl, N., & Mlakar, I. (2022). Scoping review on the multimodal classification of depression and experimental study on existing multimodal models. *Diagnostics*, *12*(11), 2683.
5. Zhang, Z., Lin, W., Liu, M., & Mahmoud, M. (2020, November). Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)* (pp. 344-350). IEEE.
6. Zafar, F., Alam, L. F., Vivas, R. R., Wang, J., Whei, S. J., Mehmood, S., ... & Nazir, Z. (2024). The role of artificial intelligence in identifying depression and anxiety: a comprehensive literature review. *Cureus*, *16*(3), e56472.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

7. Barua, P. D., Vicnesh, J., Lih, O. S., Palmer, E. E., Yamakawa, T., Kobayashi, M., & Acharya, U. R. (2024). Artificial intelligence assisted tools for the detection of anxiety and depression leading to suicidal ideation in adolescents: a review. *Cognitive Neurodynamics*, *18*(1), 1-22.
8. Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., ... & Eskofier, B. M. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, *3*(1), 66.
9. Senaratne, H., Oviatt, S., Ellis, K., & Melvin, G. (2022). A critical review of multimodal-multisensor analytics for anxiety assessment. *ACM Transactions on Computing for Healthcare*, *3*(4), 1-42.
10. Mohammad, F., & Al Mansoor, K. M. (2024). MDD: a unified multimodal deep learning approach for depression diagnosis based on text and audio speech. *Computers, Materials & Continua*, *81*(3), 4125-4147.
11. Mamidisetti, S., & Reddy, M. (2022). Multimodal depression detection using audio, visual and textual cues: A survey. *NeuroQuantology*, *20*(4), 325-336.
12. Pathirana, A., Rajakaruna, D. K., Kasthurirathna, D., Atukorale, A., Aththidiye, R., & Yatipansalawa, M. (2024). A reinforcement learning-based approach for promoting mental health using multimodal emotion recognition. *Journal of Future Artificial Intelligence and Technologies*, *1*(2), 124-142.
13. Pavlopoulos, A., Rachiotis, T., & Maglogiannis, I. (2024). An overview of tools and technologies for anxiety and depression management using AI. *Applied Sciences*, *14*(19), 9068.
14. Kumar, G., Das, T., & Singh, K. (2024). Early detection of depression through facial expression recognition and electroencephalogram-based artificial intelligence-assisted graphical user interface. *Neural Computing and Applications*, *36*(12), 6937-6954.
15. Wang, C., Liang, L., Liu, X., Lu, Y., Shen, J., Luo, H., & Xie, W. (2021, November). Multimodal fusion diagnosis of depression and anxiety based on face video. In *2021 IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE)* (pp. 1-7). IEEE.
16. Takon, A. (2022). Advanced AI Techniques for Safety and Risk Evaluation in High-Hazard Engineering Systems. *International Journal of Technology, Management and Humanities*, *8*(04), 97-109.
17. Goel, N. Privacy Risks and Protection in the Digital World of IoT. *Panamerican Mathematical Journal*, *33*(1), 2023.
18. Takon, A. (2020). Adaptive Pipeline Monitoring Using Unsupervised Anomaly Detection. *International Journal of Technology, Management and Humanities*, *6*(03-04), 93-106.
19. Singh, S. S. (2022). Accessibility and Universal Design in Transportation Infrastructure. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *14*(04), 210-214.

Pages: 90-112
Volume-I, Issue-IV, 2024
https://evjai.com/index.php/evjai

20. Takon, A. (2021). AI Safety Systems and Risk Analytics for High-Hazard Engineering Systems. *Multidisciplinary Innovations & Research Analysis*, *2*(2), 1-20.
21. Kola, J. N. (2023). Quantifying Revenue Impact of Enterprise Analytics: A Revenue Attribution Framework for Business Intelligence Systems.
22. Takon, A. (2023). Machine Learning (ML)–Based Cyber Threat Modelling for Industrial Control Systems in critical Infrastructure. *International Journal of Technology, Management and Humanities*, *9*(02), 94-108.
23. Singh, S. S. (2023). Code Compliance Challenges in High-Stakes Infrastructure Projects. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *15*(01), 213-221.
24. Kola, J. N. (2023). Measuring the Business Value of Analytics-Driven Decisions: A Decision Impact Attribution Framework for Enterprise Environments.
25. Singh, S. S. (2023). Architectural Identity in Transit Infrastructure: Branding vs Functionality. *Multidisciplinary Innovations & Research Analysis*, *4*(2), 1-12.
26. Singh, S. S. (2023). Human-Centered Design in Underground Transit Environments. *Multidisciplinary Innovations & Research Analysis*, *4*(3), 1-20.
27. Takon, A. (2024). Data-Driven Threat Intelligence for Energy and Critical Asset Management. *International Journal of Technology, Management and Humanities*, *10*(04), 253-266.
28. Kola, J. N. Longitudinal Cohort Intelligence for Self-Insured Employer Groups: A Predictive Framework for Healthcare Cost Trajectory Modeling and Proactive Risk Intervention.
29. Adepoju, S. A., & Adepoju, M. A. (2024). From Portals to Case Graphs: A Reference Architecture and Benchmark for Safety Investigation Operations with Agentic Orchestration.
30. Takon, A. (2024). Data Science Approaches to Asset Integrity Management in Offshore and Onshore Oil and Gas Operations. *Multidisciplinary Innovations & Research Analysis*, *5*(2), 17-31.
31. Goel, N. Zero Trust Architecture: A Revolutionary Approach to Cybersecurity.
32. Kola, J. N. (2011). An Integrated Framework for Data Mining and Distributed Database Optimization in Resource-Constrained Network Environments. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *2*(02), 82-86.
33. Ravikumar, V. (2014). Fair and optimal resource allocation in wireless sensor networks.
34. Naidu, K. J. (2014). Secure OLAP Reporting Architectures: Integrating Role-based Access Control and Query Execution Plan Optimization for Enterprise Analytical Environments. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *5*(02), 155-159.