FMDB *Transactions on Sustainable Health Science Letters*

Comparative Analysis of Multimodal Data Fusion Techniques for Automated Mental Health Monitoring

Aswana Lal[1,*], Chettiyar Vani Vivekanand[2]

[1,2]Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India.
aswanalal@rajalakshmi.edu.in[1], chettiyarvanivivekanand@rajalakshmi.edu.in[2]

Abstract: According to some major mental diseases, MDD is estimated to impact around 332 million people globally, and it contributes to about 727,000 cases of suicide annually. HAMD-17 and PHQ-9 are examples of traditional methods of clinical assessment of patients, which involve subjective, episodic assessment and can lead to inter-rater differences, memory recall bias, and cultural challenges to disclosure. This study combines speech acoustics, natural language processing, facial dynamics, physiological biosignals, and behavioral biomarkers gathered via passive digital phenotyping to provide an organized analysis of the multimodal computational methods for automatic depression assessment. Taxonomies of five axes input modality, feature extraction technique, learning architecture, fusion strategy, and operational context are adopted in classifying forty papers published from 2020 to 2025. Comparing cross-modal attention fusion techniques early, late, and hybrid, quantifying cross-dataset generalization capabilities for DAIC-WOZ, D-Vlog, LMVD, and BlackDog challenges, concentrating on behavioral biomarkers like psychomotor function, sleep physiology, gait kinematics, typing rhythms, and passive smartphone measurements, and delineating the future directions for explainable artificial intelligence, demographically biased models, and federated architectures for privacy preservation. The main aspects of data extraction and classification are always expressed analytically. While generalization gaps of 15% to 22% across datasets remain a hurdle for clinical application, multimodal hybrid fusion models outperform unimodal models, achieving F1 scores up to 0.84.

Keywords: Multimodal Depression Detection; Speech-Based Assessment; Behavioral Biomarkers; Affective Computing; Digital Phenotyping; Federated Learning; Mental Health Monitoring.

Received on: 01/06/2025, **Revised on:** 06/08/2025, **Accepted on:** 19/10/2025, **Published on:** 07/05/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSHSL

DOI: https://doi.org/10.69888/FTSHSL.2026.000639

Cite as: A. Lal and C. V. Vivekanand, "Comparative Analysis of Multimodal Data Fusion Techniques for Automated Mental Health Monitoring," *FMDB Transactions on Sustainable Health Science Letters*, vol. 4, no. 2, pp. 141–153, 2026.

1. Introduction

Mental disorders are among the most significant concerns affecting modern society; however, they continue to receive much less attention compared to other physical ailments. Among the most prevalent disorders, Major Depressive Disorder (MDD) is a major form of disability and is prevalent among more than 300 million people across the globe. Figure 1 shows the global depression statistics, contributing significantly to deaths by suicide each year [1]. This condition not only affects individuals but also families, institutions, and countries. The availability of experts in this field remains limited, especially in developing

[*]Corresponding author.

countries, where social stigma discourages people from seeking the help they need [2]. As a result, many cases go unnoticed or untreated. Structured interviews and various tests, such as the PHQ-9 and HAMD-17, which rely primarily on subjective responses and occasional observation, form the basis of traditional diagnosis [3]. Due to recall bias and inaccurate reporting, this approach often fails to detect symptoms at an earlier stage. This difference underscores the need to adopt objective evaluation methods in mental health assessment. Studies on computation focused on using auditory signals to detect depression, given the disadvantages of conventional approaches to the problem.

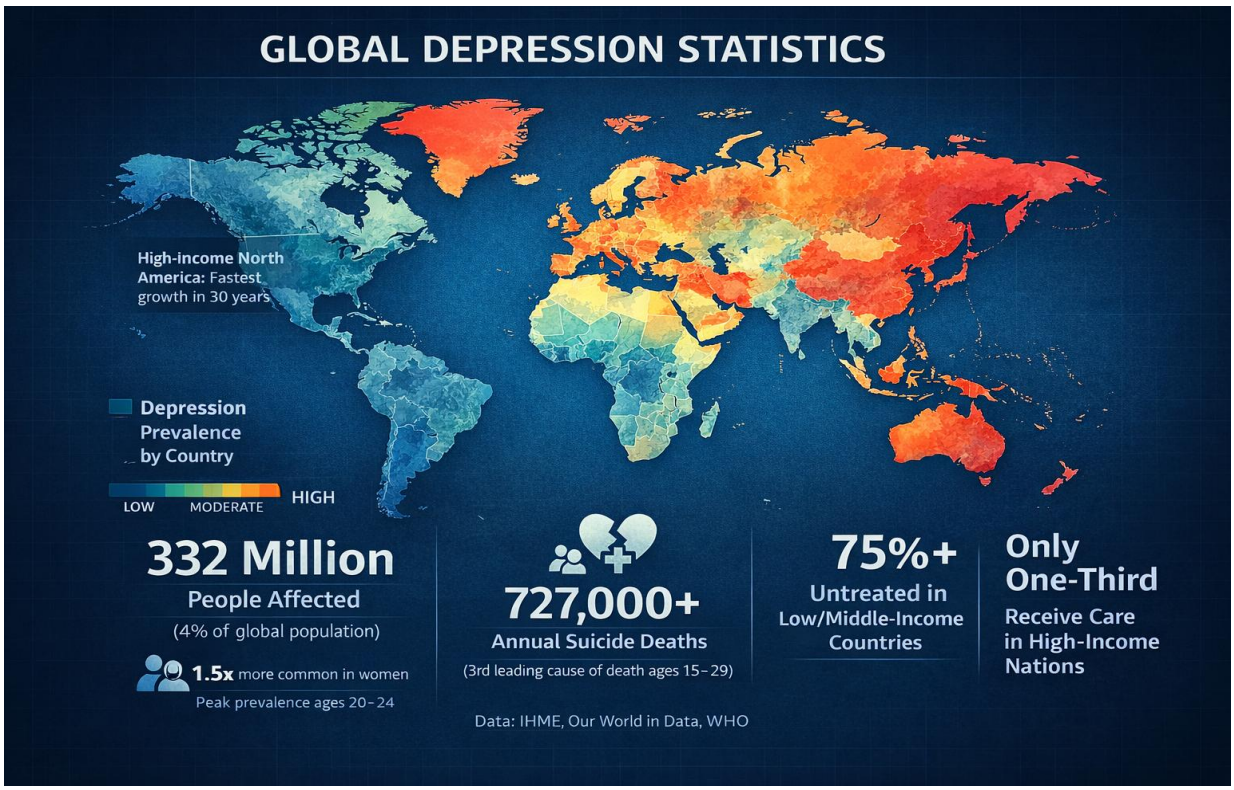

Figure 1: Shows the global depression statistics infographic

Auditory features such as pitch, loudness, and pauses have been identified in some studies as reflecting certain emotions and cognitions [4]. Machine learning models, including support vector machines, have been used for classification purposes based on these signals associated with depression [5]. The validity of speech as a reliable, non-invasive biomarker is established because depressed individuals tend to speak slowly and exhibit limited vocal variation [6]. These results thus laid the groundwork for the development of automated mental health screening tools. But given the variability in emotional expression from one individual to another, speech alone could not account for all the complexities of human behavior. Nevertheless, the study conducted during the early stages of speech processing demonstrated that physiological biomarkers could significantly improve depression screening systems. Academics Figure 2 shows the evaluation of methodologies-initiated research using text data and behavioral information as supplementary means to help understand psychological problems apart from voice analysis. The emergence of social media platforms, where individuals express themselves using everyday language, became essential for obtaining information [7]. Findings showed that despite the lack of visible signs, depression could be diagnosed based on language use, mood, and word choice [8].

This approach enabled tracking trends in mental health without involving clinicians. Moreover, the notion of "digital phenotyping" enabled continuous assessment of mental state using behavioral data, such as phone use and movement [9]. Research has also revealed that those with depression tend to show signs of lower social interaction and exercise activities [10]. Such developments indicated the importance of combining multiple sources of information for a better understanding of all symptoms of depression. Nonetheless, because each of these sources provides limited information about the individual's psychological state, relying on a single source poses challenges. The current research (Figure 3) shows that effort has been directed toward multimodal systems that incorporate modalities such as speech, text, and behavior to address the limitations of unimodal approaches. Benchmarking datasets such as DAIC WOZ has helped develop models based on clinical interviews. Likewise, standardized benchmarks for evaluating the emotion and depression detection systems have been offered by the AVEC data set [12]. Deep learning models have significantly improved their effectiveness by automatically extracting features

from raw data [14]. In speech and language processing, transformers have been successful at improving context understanding [15].

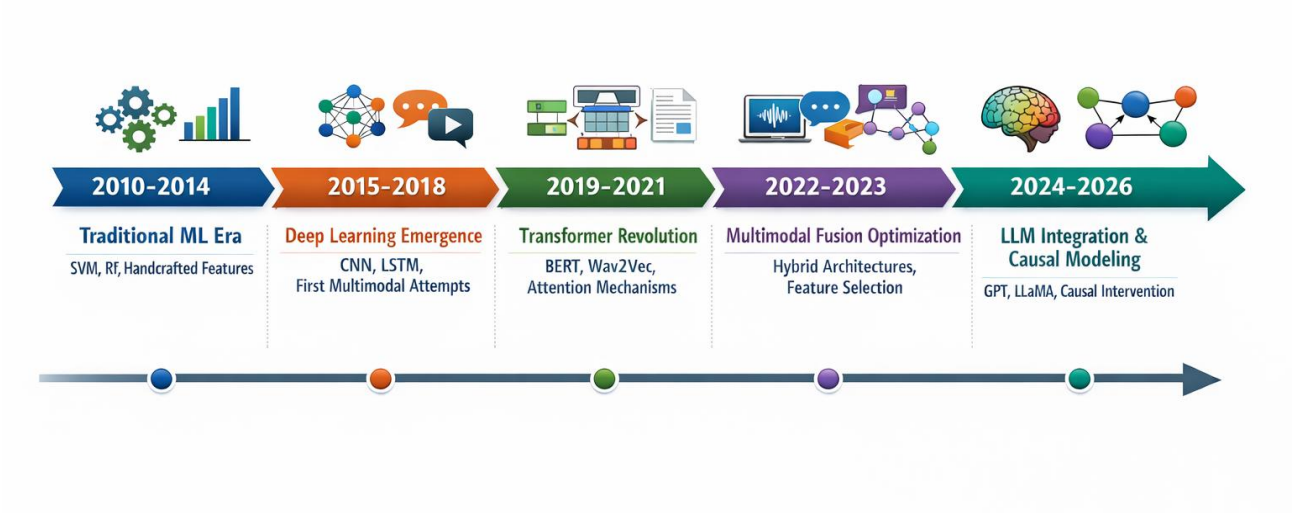

Figure 2: Evolution timeline of methodologies

Notwithstanding the advancements made in these areas, some challenges remain. Issues related to the use of personal data have been raised due to concerns about data privacy [16]. In addition, the accuracy of the predictions may be affected by the algorithms themselves, which can introduce algorithmic bias [17]. Multi-modal learning solutions have been proposed to address such challenges, leveraging multiple data sources to achieve greater accuracy [18].

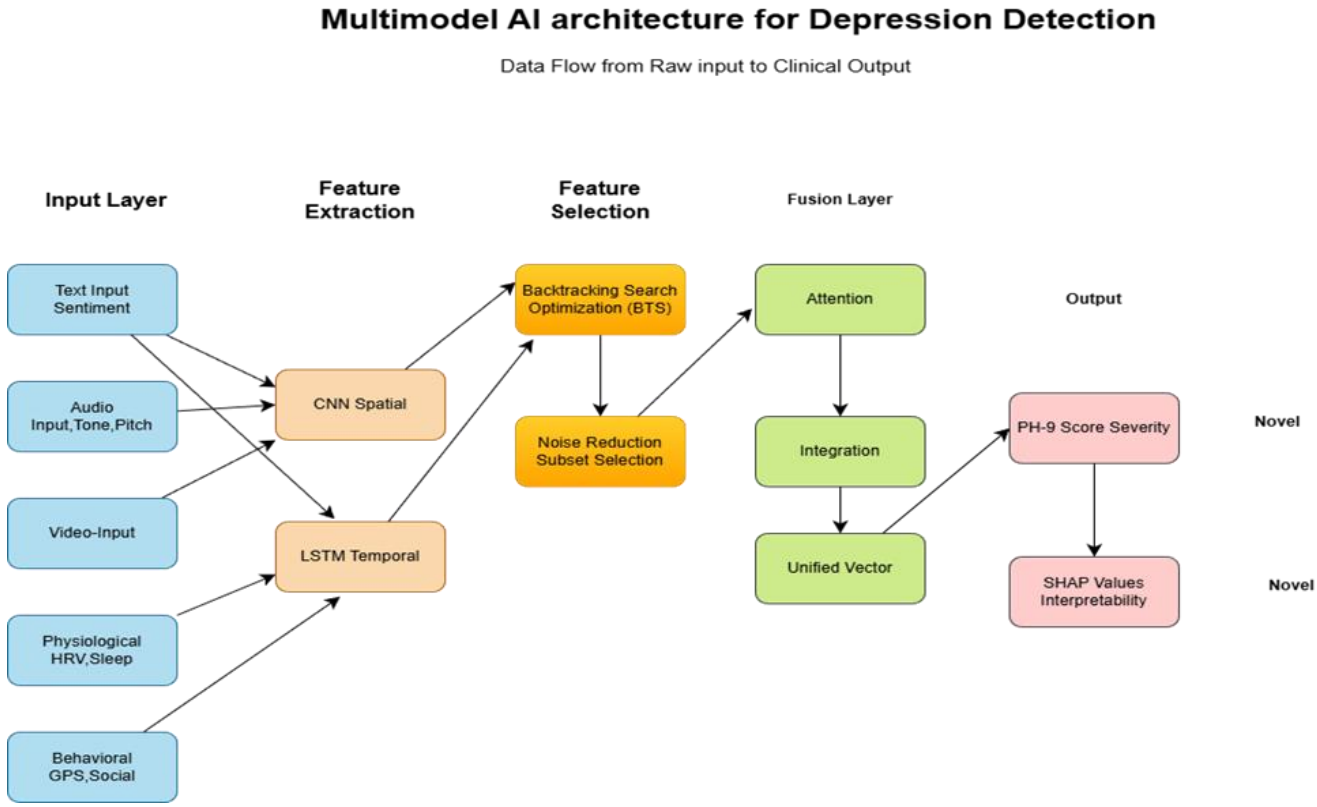

Figure 3: Technical framework with novel contributions highlighted

Although there have been remarkable developments, several unresolved problems exist in the domain of computational depression detection. Comparing and assessing the performance of different models remains difficult due to the lack of

standardized methods and evaluation schemes. Although most models demonstrate high precision in controlled studies, they are unable to generalize their performance across diverse conditions and individuals. Moreover, data handling becomes increasingly complex as multiple types of information are combined. An approach that incorporates knowledge from multiple areas of study must be used to address these issues. Therefore, this research paper aims to conduct an integrated analysis of approaches to automated depression detection. This research should aid in developing reliable and ethical methods for mental health testing by analyzing existing tools and identifying further research needs. In the long run, such innovations can lead to new methods of diagnosing depression, making it possible to start treating the illness earlier [19].

2. Background

Depression, anhedonia, cognitive impairment, sleep disturbances, psychomotor changes, and even suicidal ideas in the most serious cases are the defining features of the chronic mental health disorder known as major depressive disorder. Apart from emotional disturbances, major depression is actually a condition that involves faulty neural pathways. Several functional neuroimaging studies have revealed hypometabolism in the hippocampus, caudate nucleus, and dorsolateral prefrontal cortex, along with hypermetabolism in the amygdala and subgenual anterior cingulate cortex [11]. The errors associated with emotional regulation in acoustic and behavioral output are regulated via the hypoconnectivity between the dorsolateral prefrontal cortex and the amygdala, thereby providing objectives for computational models with biological significance [4]. Psychomotor retardation associated with moderate to severe depression results from dysfunction of basal ganglia pathways, as indicated by reduced gait velocity, increased motor reaction times, decreased facial muscle velocity, and a slower speech rate. Abnormal regulation of the hypothalamus-pituitary-adrenal axis leads to elevated cortisol levels in the bloodstream, which impairs the hippocampus' regenerative capacity. Consequently, it causes cognitive impairments that appear as greater lexical scarcity, ideation repetition, and reduced semantic variation in language generation. This is why these molecular processes scientifically explain the selection of objective markers in the form of behavioral and vocal signs, as they represent direct symptoms of neuropathology rather than coincidental associations. Hence, the medical importance of these two signal types should be attributed not just to their empirical association but also to their role in neurophysiology [5].

2.1. Clinical Assessment Instruments and Ground Truth Labeling

The technology behind the detection system is built on the technical principles of representation learning and signal processing. Audio signals are converted into a structured representation, such as a Mel spectrogram, or a condensed representation using acoustic descriptors of temporal change, spectrum energy, and frequency dynamics in speech analysis. Embedding models that preserve semantic and contextual relationships among sequences of data are used to represent textual data. Spatial-temporal modeling of facial muscle movement and micro-expressions is used to interpret the visual data stream. To ensure analytical consistency, each modality must be aligned, normalized, and noise-reduced. The preprocessing process for these modalities aims to reduce the influence of environmental factors while creating a consistent mathematical representation that preserves diagnostically relevant information.

The representation will then be used to apply a learning algorithm that will recognise the non-linear relationship between depressed symptomatology and behavioral patterns. The Hamilton Depression Rating Scale (seventeen-item version - HAMD-17), the Patient Health Questionnaire (nine-item version - PHQ-9), the Beck Depression Inventory second edition (BDI-II), and the Structured Clinical Interview for DSM Disorders (SCID) are some of the most commonly used methods in the field of clinical depression diagnosis. They are utilized as gold standard labels in many studies on computing depression. All these scales use ordinal labeling according to their severity for multiple symptom categories, including suicidal ideation, anhedonia, sleep disturbance, psychomotor agitation/retardation, depressed mood, and concentration problems. Scores of HAMD-17 above 24 imply severe depression, while scores above 15 on the PHQ-9 scale imply moderate-severe depression. All these scales have inter-rater reliability ranging from 0.70 to 0.85, which introduces noise into the label values.

2.2. Signal Processing and Feature Extraction

2.2.1. Mel-Frequency Cepstral Coefficient Extraction

The Mel-Frequency Cepstral Coefficients (MFCCs) have been widely used for auditory representation in spoken depression detection. The successive steps in Equation 1 that are required to extract the MFCC are as follows: Pre-emphasis is a filter that compensates for the spectral rolloff in speech production:

$$y[n] = x[n] - a * x[n-1], \quad a \text{ in } [0.95, 0.97] \tag{1}$$

The Short-time Fourier Transform with Hamming window w(n) of length N applied to overlapping frames is represented in Equation 2:

$$X(m,k) = SUM_{n=0}^{N-1} x[n+mH] * w(n) * exp(-j2*pi*k*n/N) \quad (2)$$

Equation 3 shows Mel-scale frequency mapping that approximates human auditory perception:

$$mel(f) = 2595 * log10(1 + f/700) \quad (3)$$

Equation 4 Application of M triangular Mel filterbanks to the power spectrum:

$$S(m,l) = SUM_{k=0}^{N/2} |X(m,k)|^2 * H_l(k), \quad l = 1,...,M \quad (4)$$

Discrete Cosine Transform to decorrelate filterbank outputs into cepstral coefficients represented by Equation 5:

$$c(m,i) = SUM_{l=1}^{M} log(S(m,l)) * cos[pi*i*(l-0.5)/M], i = 1,...,K \quad (5)$$

In each case, the number of cepstral coefficients (K) is kept between 13 and 40, and delta and double-delta operations are applied for time-domain modeling. As a result, researchers can obtain feature vectors with dimensions ranging from 39 to 120, depending on the configuration used.

2.2.2. Prosodic Feature Formulations

Autocorrelation techniques are used to extract the fundamental frequency (F0) contour. Over the entire speech duration, depression-relevant statistics summaries are calculated in Equation 6:

$$F0\_stats = \{mean(F0), std(F0), min(F0), max(F0), range(F0), skskewness(F0)\} \quad (6)$$

Jitter, measuring period-to-period F0 variation, and shimmer, measuring amplitude variation, are computed in Equations 7 and 8:

$$Jitter(\%) = [(1/(N-1))*SUM|T_i - T_{i+1}|] / [(1/N) * SUM(T_i)] * 100 \quad (7)$$

$$Shimmer(dB) = 20*log10((1/(N-1))*SUM|A_i- A_{i+1}|)/((1/N)*SUM(A_i)) \quad (8)$$

In voice activity detection pre-processing, speaking rate can be measured in syllables per second. The ratio of the total duration of silence to the total duration of the recording session is called the pause ratio and is higher in depressed speech than in normal speech.

2.2.3. Transformer-Based Self-Supervised Representation Learning

Waveform representations are learned using self-supervised methods such as HuBERT and wav2vec 2.0, which employ a context-aware transformer encoder after a convolutional feature extractor. Equation 9 shows the latent representation generated by the convolutional feature extractor for the input waveform x with a length of T:

$$Z = f\_enc(x) = CNN(x), Z \text{ in } R^{T' x D} \quad (9)$$

Equation 10, the sequence is subjected to multi-head scaled dot-product self-attention by the transformer encoder:

$$Attention(Q,K,V) = softmax(Q*K^T / sqrt(d\_k)) * V \quad (10)$$

To predict the degree of depression, a specialized classification head is employed on the average value from the encoded vector: y-hat = sigma (W * mean(Z) + b), wherein “sigma” denotes sigmoid activation for binary classification and softmax for predicting the degree.

2.2.4. Behavioral Biomarker Quantification

Location entropy, a metric that captures the spatial range and variety of individual movement, is used to summarise passive smartphone mobility data in Equation 11:

$$H\_loc = - SUM\_k \ p(k) * log(p(k)) \quad (11)$$

Where the proportion of the time taken during the observations within location clusters k is represented by p(k). A much smaller location entropy can be observed for those suffering from depression, implying less spatial movement compared to their baseline. The use of the z-score deviation calculation is applied for individual baseline normalization:

$$D(t) = (z(t) - mu\_personal) / sigma\_personal \quad (12)$$

In Equation 11, where p(k) represents the proportion of total observation time spent in the location cluster k. The location entropy value is significantly smaller in patients with depression, suggesting reduced spatial mobility relative to their individual baselines. The calculation of Z-score deviation helps in obtaining personal baseline normalization. In contemporary depression diagnosis systems, machine learning architectures that can handle high-dimensional, time-varying data are applied. Although they initially appeared viable, statistical classifiers struggled to handle complex sequential relationships. Hierarchical representation learning and temporal pattern learning using deep learning techniques, such as convolutional and recurrent neural networks, can address these problems. The application of attention mechanisms can also hone the process by focusing on the most important segments of voice, text, or video that significantly impact the results. The basic issues of class imbalance, demographic bias, cross-cultural differences, dataset availability, and interpretability in clinical scenarios remain the same despite technological advances. The primary aim of research on multimodal depression detection is to develop reliable, scalable, and ethical computer systems that can enhance psychiatric knowledge by objectively quantifying behavior.

3. Taxonomy of Multimodal Depression Detection Systems

3.1. Input Modality

The different types of systems are classified based on the nature of the input signal they study, using the main taxonomic dimension. Systems based on speech prosodic, spectral, and voice quality features, in relation to the biomarker of depression, using acoustics captured during vocalization. Systems based on text-capture information related to semantic negativity, syntactic complexity, sentiment polarity, and lexicon richness from either written or transcribed language. The systems are based on visual capture of movement energy, head posture kinematics, gaze trajectories, and action unit activation from facial video feeds. Examples of systems based on physiological input include EEG, ECG, heart rate variability captured from PPG, and electrodermal activity. Finally, behavioral systems use passive tracking from wearable or smartphone sensors to monitor movement, sleep, mobility, and social behavior.

3.2. Feature Representation Strategy

Four eras of the development of automation and abstraction are depicted through feature representation approaches. In the first generation, the process involves creating manual feature representations that require domain knowledge and are easily influenced by recording conditions. Such manual features include statistics on MFCC, pitch, pauses, and lexical frequency. The next generation includes engineering feature representations such as the extended Geneva Minimalistic Acoustic Parameter Set, which contains 88 standardized acoustic features, the Linguistic Inquiry and Word Count approach used in psychological text analysis, and Action Units. Third-generation deep learning representations employ embeddings obtained using transformer architectures such as BERT and RoBERTa for text input, convolutional neural networks for mel-spectrogram inputs, and bidirectional long short-term memory networks for acoustic features. Fourth-generation representations employ self-supervised learning and perform better through transfer learning based on statistics of speech and language. They are pre-trained on very large unlabelled datasets and then fine-tuned on a smaller labeled dataset related to depression [6].

3.3. Learning Architecture

Learning architectures vary from high-capacity deep learning architectures to more interpretable shallow models. In conjunction with carefully curated feature sets, radial basis function kernel-based support vector machines prove highly efficient learners for data sets comprising fewer than 200 participants, achieving F1 scores of 0.71-0.76. Variance reduction can improve the robustness of ensembling algorithms such as random forests and gradient-boosted decision trees. Convolutional neural network spectrogram classifiers, long short-term memory networks, bidirectional LSTM temporal sequence models, temporal convolutional networks, and transformer encoders with cross-modal attention are among the initial attempts at deep learning architectures. The highest F1 scores for the text modality in the 0.87-0.91 range, achieved using a large-language-model-enhanced architecture on depression-related content, come at a considerable computational cost [7].

3.4. Multimodal Fusion Strategy

The fusion of modality-specific representations depends on the methods used. Early fusion combines features of all input modalities into one single feature vector F_joint =[f_audio||f_text||f_video] belonging to R^(D_a+D_t+D_v) before

classification. Although early fusion ensures that features from different modalities interact at a certain level, it has weaknesses such as noise propagation from weaker modalities and dimensional imbalance. In late fusion, features of each input modality are individually classified by specialized classifiers and then averaged to arrive at the final prediction (P_final=SUM_m(w_m*P_m(x_m)). While keeping modality-specific model optimizations intact, cross-modal interplay at the inference stage cannot be achieved. To dynamically allocate modality importance, hybrid and attention-based fusion techniques use cross-modal attention transformer networks, where A_ij = SoftMax ((Q_i * K_j^T)/sqrt(d)). This allows for inter-modality attention between textual, auditory, and visual features during inference. Although it consistently beats benchmarks, this technique requires significantly more CPU power and time synchronization of inputs [8].

3.5. Deployment

The tradeoffs between ecological validity and diagnostic accuracy are illustrated across five implementation environments. The clinical interviews approach emphasizes diagnostic accuracy over scalability, using audio recordings of structured psychiatric consultations that are coded using well-known clinical rating scales. The social media screening tool provides broad population coverage at the expense of clinical accuracy and ethical considerations by analyzing publicly available content from social networking sites such as Reddit and Twitter using natural language processing techniques. Anonymous and accessible screening is enabled by artificial intelligence platforms that embed computational components into the conversation process between humans and computer systems. Laboratory biosignal analysis machines achieve high signal accuracy but are limited in practical application due to hardware and knowledge prerequisites.

4. Comparative Analysis

4.1. Dataset Overview and Limitations

The primary benchmark datasets employed in multimodal depression diagnosis studies are listed in Table 1. The major constraint is that the datasets involve fewer than 300 subjects, which makes it difficult to train highly capable deep learning algorithms, as there is always a risk of overfitting with small datasets. For many benchmarks, demographic constraints limit the generalizability of results, as they involve only subjects from educated, English-speaking Western countries.

Table 1: Lists the datasets used in multimodal depression detection studies

Dataset	Year	Modalities	Subjects	Language	Label	Environment	Primary Limitation
DAIC-WOZ	2014	Audio + Video + Text	189	English	PHQ-8	Interview	Small sample; Western demographic only
D-Vlog	2022	Audio + Video + Text	299	Korean	BDI-II	Vlog	Cultural and linguistic bias; Korean only
LMVD	2023	Audio + Video + Text	240	Mandarin	PHQ-9	Interview	Language limited; single recording site
BlackDog	2018	Audio + Video	255	English	SCID	Clinical	Controlled laboratory setting only
Pitt	2019	Audio only	104	English	SCID	Clinical	Audio unimodal; small sample
PDCH	2025	Speech + Text	100	Mandarin	HAMD -17	Clinical consult	Single hospital; limited demographic scope

4.2. Fusion Strategy Performance Comparison

The performance of different fusion methods, architectures, and data sets is quantitatively analyzed in Table 2 and Figure 4. Under similar data set conditions, fusion methods using attention mechanisms consistently outperform both early and late fusion by four to eight F1 points.

Table 2: Comparative performance of multimodal depression detection systems across modalities, fusion strategies and architectures

Study	Modalities	Fusion Strategy	Architecture	F1 Score	Benchmark Dataset
Jin et al. [1]	Text + Speech + Facial	Hybrid Attention	LLM + CNN + Tri-modal	0.89	DAIC-WOZ

De Almeida et al. [3]	Audio-Visual + Text	Stacked Ensemble	BiLSTM + DCNN + TCN	0.787	D-Vlog / EDAIC
Lu and Fu [13]	Audio + Text	Early Fusion	CNN + BiLSTM + Attention	0.79	DAIC-WOZ
Zhang et al. [12]	Audio + Video + Text	Hybrid Fusion	CNN + TF-IDF + MFCC	0.82	DAIC-WOZ
Zhou et al. [14]	Speech + Facial	Late Fusion	ML + Prosodic + AU	0.74	Clinical sample
Chiong et al. [17]	Text only	Unimodal	Random Forest + TF-IDF	0.84	Twitter corpus
Wang et al. [16]	Audio only	Unimodal	3D-CNN Spectrogram	0.71	DAIC-WOZ
Asare et al. [18]	Behavioral only	Unimodal	Random Forest + Smartphone	0.81	Smartphone corpus

The performance of the hierarchical approach to fusion is consistent across the examined literature. A baseline for a unimodal approach ranges from 0.71 to 0.84, depending on the modality; an approach with hybrid attention-based fusion gives an F1 of about 0.84; an approach with early fusion gives an F1 of about 0.79; an approach with late fusion gives an F1 of about 0.74.

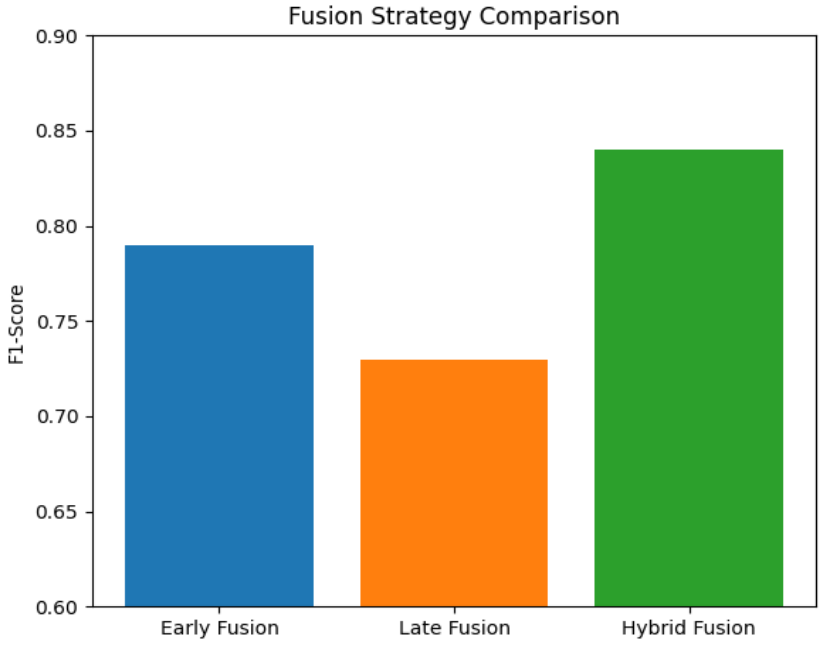

Figure 4: Performance comparison of early, late, and hybrid multimodal fusion using F1-score

For passive data collection using smartphones, the behavioral unimodal approach developed by Asare et al. [18] yields an F1 score of 0.81, indicating that a passive behavioral sensing approach alone provides a significant and independent diagnostic signal.

4.3. Cross-Dataset Generalization Analysis

Cross-dataset generalization (Figure 5) poses the primary unsolved challenge in computational diagnosis of depression. Performance reductions of 10 to 22 percent have been reported after training a model using one benchmarking dataset and testing it against another benchmarking corpus. Models trained on DAIC-WOZ experience a 18% decrease in F1 on D-Vlog and a 22% decrease on LMVD, owing to disparities in language, interviewing style, cultural and emotional expression guidelines, and acoustic recording environment. Models trained through LMVD show a reduction in performance by 20 to 21 percent when benchmarked against other English-speaking corpora, indicating that cross-linguistic generalization is the most difficult aspect of the challenge. Such reductions could be attributed to various factors, such as differences in each dataset's acoustic recording environment, incompatible clinical severity ratings, diverse demographic representations, and culturally informed emotional expression styles.

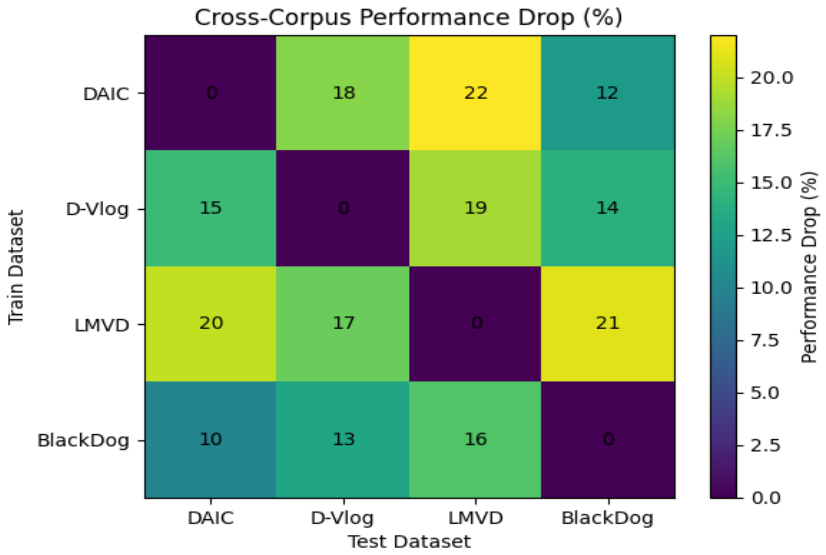

Figure 5: Cross-dataset performance drops the generalization limit across multimodal models

4.4. Architecture Complexity

Table 3 presents a structured comparison of architectural classes across key operational dimensions relevant to clinical deployment decisions.

Table 3: Architectural analysis across key clinical deployment dimensions

Architecture	Typical F1	Min. Dataset Size	Interpretability	Inference Speed	Cross-Cultural	Data Need
SVM / Logistic Regression	0.65 to 0.75	< 200 subjects	High	Fast	Moderate	Low
Random Forest Ensemble	0.72 to 0.80	< 300 subjects	Moderate	Fast	Moderate	Low
CNN (Spectrogram)	0.70 to 0.78	300 or more	Low	Moderate	Limited	Medium
BiLSTM / LSTM	0.74 to 0.82	300 to 500	Low	Moderate	Limited	Medium
CNN-LSTM Hybrid	0.78 to 0.86	500 or more	Very Low	Slow	Limited	High
Transformer Attention	0.82 to 0.89	1000 or more	Very Low	Slow	Good	Very High
LLM-Enhanced Multimodal	0.87 to 0.92	Very large	Very Low	Very Slow	Good	Extreme

4.5. Behavioral Versus Physiological Modality

The tradeoff between the physiological objectivity of the data and the scalability of behavioral data cannot be avoided. Since they can be measured with an accuracy of 85% to 92%, electroencephalographic methods exhibit the highest internal validity in laboratory settings. Nonetheless, their limited acquisition conditions, hardware requirements, and skill levels disqualify them for use at a population scale. There is a physiological data type that possesses high longitudinal capability and a good balance between scalability and objectivity (F1 between 0.72 and 0.78), namely heart rate variability derived from photoplethysmography on smartwatch measurements. Behavioral digital phenotyping is built upon sacrificing its physiological foundation in exchange for higher scalability and longitudinality. Federated multi-device sensing is one approach that can ensure physiological objectivity and scalability. In this context, edge computing devices such as smartphones, smartwatches, and microphones collaborate to generate behavioral markers for depression without centralizing sensitive data. The federated multi-device sensing gradient averaging rule expression is shown in equation 13 as follows:

$$theta_{t+1} = theta_t - eta * (1/K) * SUM_k GRAD L_k(theta_t; D_k) \quad (13)$$

Where θ denotes the global learning rate, L_k refers to the local loss function on device k using local data D_k, and K indicates the total number of devices taking part. This formulation enables improving the model without exposing individual behavioral data to central storage.

5. Open Challenges and Research Gaps

5.1. Dataset Scale, Diversity, and Representativeness

The first major impediment to research progress is limited data. Typically, the standard dataset size is 100-300 clinically proven participants, which is insufficient to train powerful deep neural networks without the risk of overfitting. English-speaking, Westernly educated people account for up to 70% of currently used benchmark samples, thereby introducing an important demographic bias that undermines the ability to generalize to other deployment contexts globally. Cases with a more severe course of depression and especially comorbidity can be significantly underrepresented in severity distributions, since they are always focused on relatively light or moderate depressive states. As opposed to actual application environments, with the noise level exceeding 40 dB SNR, considerable variation of microphone quality between consumer devices, and unscripted discourse, controlled recording environments, i.e., quiet interview settings with standardized microphones and conversation protocols, create substantially different signal distributions. This is why performance rates of 68 to 78 percent are typically derived from laboratory accuracy of 85 to 92 percent.

5.2. Multimodal Synchronization and Missing Modality

The multimodal systems encounter initial signal alignment issues due to inherent disparities in the signal sampling grids. Before carrying out the joint feature extraction, explicit alignment in terms of time needs to take place between transcriptional data sampled in terms of individual utterances, video streams sampled at 25-60 frames per second, and audio streams sampled at 16 to 44kHz. Misalignment of more than 200 milliseconds tends to weaken the robustness of cross-modal inference, leading to inaccurate analysis of voice prosody and facial expression timing. Missing modality events are common during real-world deployments due to factors such as camera obstruction, insufficient lighting, limited bandwidth and user privacy concerns. There has been no reported instance of imputation for modality event dropouts, yet accuracy levels tend to decrease by 15-20 percent.

5.3. Clinical Validation Gap

The ability of machine learning models to make their decisions interpretable to clinicians in terms they can question is as crucial as their performance when considering the clinical utility of computational methods for depression detection. The primary obstacle to clinical implementation is the inability of existing deep learning models to explain how their predictions were made, as they return probability estimates or severity classifications without any mechanistic interpretation. If there were an application of the population-based screening method for the identification of depression, at least a clinically meaningful proportion of the depressed people would suffer from delayed treatment due to a false-negative rate higher than 10% to 15%. Post-hoc assessment of feature significance is achieved using explainable AI tools such as gradient-weighted class activation mapping for spectrograms of convolutional CNNs, LIME for text-based predictions, and SHAP for acoustic characteristics. Nevertheless, clinical verification against clinicians' observations is needed to validate these hypotheses, especially to determine whether the identified computational factors align with established psychological disorders.

5.4. Temporal Modeling and Longitudinal Depression Dynamics

Depression has been demonstrated to have a recurrence rate ranging from 50% to 80% during the first five years after its primary onset, making it a temporal and episodic disorder. Current computational studies have used recordings of 10 to 20 minutes, providing a static assessment of a longitudinal condition, which cannot capture dynamics related to remission and/or relapse. Fewer than 5% of the reviewed studies have collected data on monitoring sessions exceeding 3 months, and even fewer have reported monitoring sessions lasting more than 1 year. Population-normalized thresholds cannot be used to predict individual outcomes because inter-individual variability exceeds intra-individual differences in baseline vocal energy and speech rate. One of the critical knowledge gaps that makes detecting relapses difficult is the characterization of behavioral markers' evolution through treatment stages, drug changes, and psychosocial stressors.

5.5. Ethical, Privacy, and Regulatory Challenges

The biometric data being collected is highly sensitive, including facial features and the speaker’s voice. The biometric data being used constitutes legal unique identifiers in some regulatory regimes, which could be re-identified and abused. Based on epidemiological studies, more than 60 percent of the potential users of the technology claim that they feel reluctant to share

their mental health data with remote servers. False positive results of identifying depression in the population might pose serious iatrogenic risks in relation to employment, insurance, and education screenings. Despite known differences in algorithm performance across age, gender, race, and socio-economic status in affective computing, such distinctions are seldom assessed specifically in the field of depression recognition. Automated health care solutions are regulated by diverse and even conflicting regulatory frameworks worldwide. Thus, under the EU Artificial Intelligence Act, automated mental health technologies are deemed high risk and thus fall under the regime of conformity assessments. At the same time, the FDA has not developed any clearance processes for computational psychiatry applications. Governance practices such as transparency in consent, data minimization standards, independent auditing, and algorithmic fairness evaluation are essential for ethical application.

6. Future Research

6.1. Advanced Cross-Modal Attention Architectures

In contrast to treating cross-modal interactions as a basic principle underlying the architecture design process for the entire representation learning task, today's state-of-the-art multimodal models predominantly treat it as a post-learning fusion process. To enable auditory, textual, and visual representations to constrain one another during encoding, future multimodal architectures should be designed so that end-to-end cross-modal attention is applied from the feature-extraction level onward. The model will be able to detect covariances that are clinically relevant, such as the additional relevance of diminished pitch variation due to the simultaneous presence of flattened facial affect along with first-person negative semantics, due to the presence of cross-modal transformers that have text query vectors that pay attention to audio and video key value pairs. In the absence of labeled multimodal depression training sets, it is possible to use contrastive multimodal pre-training on large unlabeled behavioral data, following the precedent set by general vision-language research.

6.2. Behavioral Digital Phenotyping at Scale

Passive behavioral phenotyping via consumer devices employed in natural settings represents the most compelling route towards scalable longitudinal monitoring of depression. To develop consistent depression biomarkers that operate continuously over clinically meaningful periods, future platforms must include behavioral data captured by smartphones, wearable devices, and ambient-sensing devices. Federated learning frameworks, which provide a solution to the critical privacy constraint, while at the same time offering a source of diverse population-level behavioral data for model building, as shown in Equation 13, are an important component of future systems. Personalized baseline normalization, which accounts for inter-person variability that hinders population-based thresholding approaches, as demonstrated in Equation 12, is an essential factor that enables the detection of abnormal behaviors based on deviations from individual rather than population-level baselines.

6.3. Longitudinal Architectures and Personalized Relapse Detection

Temporal encoding models based on attention mechanisms, which incorporate daily activity data recorded over 60–90-day observation periods, can be used to track an individual's depression history, detect preconditions for future relapse, and evaluate the quality of therapy received. An architecture of this nature requires ongoing recording of behavioral data under changing conditions, employing methods of domain adaptation and transfer learning to ensure the model operates stably across varying conditions involving changes in equipment usage, geographical location, and living context. The goal shifts from classifying the depressive episode to predicting it.

7. Conclusion

By synthesizing more than 40 papers published between 2020 and 2025 across speech acoustics, natural language processing, facial dynamics, physiological biosignals, and behavioral digital phenotyping, this review article offers an exhaustive hierarchical survey of the state-of-the-art computational techniques used in automated depression diagnosis based on multimodal data. To facilitate an analytical framework for the comparative study of the papers, a five-factor taxonomy of the input modality, feature representation method, learning paradigm, fusion strategy, and operational environment has been established. For the sake of technical replicability and didactic purposes, the mathematical formulations of the extraction of MFCC, prosody, transformer, entropy, deviation, and federated gradients were included. Behavioral digital phenotyping is a scalable, ecologically valid complementary technology alongside acoustic and facial technologies, while multimodal, hybrid, attention-based fusion techniques are the most effective for detection, with F1 scores ranging from 0.84 to 0.92 in a literature review.

The main barrier to adopting this approach in clinical practice lies in the generalization gaps between benchmarking datasets and the population in actual clinical conditions. These gaps range from 15% to 22% and stem from persistent disparities in the distribution of demographics, language, and environment between deployment populations and benchmark datasets. The

progress of computational mental health assessment is getting nearer to the point where it will be possible to objectively track any changes in behavior associated with depression among populations continuously and in a culturally sensitive fashion. It is essential to work together across disciplines in a sustained manner to make such an approach feasible while maintaining clinical safety, algorithmic justice, and data privacy. Designed using rigorous methods and ethical frameworks, such AI systems for depression assessment provide a promising avenue to ensure that most of those affected by depression and currently lacking treatment receive effective psychological support.

Acknowledgment: This literature review is pure research based only on the analysis of past research. There is no need to acknowledge anyone other than the above-mentioned authors, laboratory assistants, or external collaborators, as no one else assisted with the research.

Data Availability Statement: The Survey does not have any original datasets created specifically for this research; rather, it is based on a comprehensive analysis of existing sources. All statistics on performance and assessment metrics cited in this research are from prior literature and are acknowledged in the bibliography. The publicly available datasets used in this survey, such as DAIC-WOZ, D-Vlog, LMVD, BlackDog, and PDCH, can be accessed via the website addresses provided by the source communities. In addition to the data mentioned in the book, no additional data have been created for this survey.

Funding Statement: The authors state that this research and the preparation of the manuscript were completed independently without receiving any external financial aid, sponsorship, or grant support from funding agencies or organizations.

Conflicts of Interest Statement: The authors declare no conflict of interest. None of the authors has any financial, professional, or personal connections with any entity, object, or individual discussed within this manuscript that would compromise or bias the outcome, findings, or recommendations offered. All citations are based solely upon their scientific merit and relevance to the subject matter.

Ethics and Consent Statement: Regarding this literature review, the researchers have neither obtained nor generated any primary data involving human subjects, as this study is a review of the existing literature. Neither clinical nor behavioral data collection was performed. All data used in this study were obtained from other studies, and the data collection and ethical issues had already been addressed in the initial articles. As such, conducting this survey did not require additional ethics approvals and consents. The authors assure that all requirements for the responsible conduct of research were observed while preparing this publication.

References

1. Y. Jin, X. Chen, J. Liu, Z. Chen, J. Zhou, Y. Li, Y. Bu, and Y. Wang, "Harnessing multimodal emotion features in depression detection across gender: Integrating large language model, acoustic fusion and facial expression recognition," *Journal of Affective Disorders*, vol. 400, no. 5, p. 121207, 2026.
2. M. S. Haleem, V. Aidonis, E. I. Georga, M. Krini, M. Matsangidou, A. P. Kassianos, C. S. Pattichis, M. Rujas, L. Lopez-Perez, G. Fico, L. Pecchia, and D. I. Fotiadis, "A multimodal deep learning architecture for estimating quality of life for advanced cancer patients based on wearable devices and patient-reported outcome measures," *IEEE Journal of Biomedical and Health Informatics*, vol. 30, no. 2, pp. 1166–1177, 2026.
3. F. F. De Almeida, K. R. T. Aires, A. C. B. Soares, L. D. S. B. Neto, and R. D. M. S. Veras, "Multimodal analysis for depression recognition using stacked multilevel deep neural networks," *IEEE Access*, vol. 14, no. 1, pp. 10033–10052, 2026.
4. E. Jordan, R. Terrisse, V. Lucarini, M. Alrahabi, M. O. Krebs, J. Desclés, and C. Lemey, "Speech emotion recognition in mental health: Systematic review of voice-based applications," *JMIR Mental Health*, vol. 12, no. 9, p. e74260, 2025.
5. L. Berkemeier, W. Kamphuis, A. M. Brouwer, H. De Vries, M. Schadd, J. U. Van Baardewijk, H. Oldenhuis, R. Verdaasdonk, and L. Van Gemert-Pijnen, "Measuring affective state: Subject-dependent and-independent prediction based on longitudinal multimodal sensing," *IEEE Transactions on Affective Computing*, vol. 16, no. 2, pp. 827–843, 2024.
6. P. Cao, Y. Zhang, C. Zhang, W. Chen, Y. Liu, S. Xu, M. Xu, W. Jin, J. Xu, D. Wang, W. Wang, X. Wang, W. Wang, Y. Ren, J. Zhao, R. Li, and K. Liu, "A multimodal depression consultation dataset of speech and text with HAMD-17 assessments," *Scientific Data*, vol. 12, no. 1, p. 1577, 2025.
7. Y. Li, S. Kumbale, Y. Chen, T. Surana, E. S. Chng, and C. Guan, "Automated depression detection from text and audio: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 10, pp. 7498–7513, 2025.
8. M. Junaid, M. Ghergherehchi, and S. Lee, "Multitask deep learning for predicting Parkinson's progression and depression from multimodal time series data," *IEEE Access*, vol. 13, no. 7, pp. 147818–147841, 2025.

9. M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnykova, "Multimodal data fusion for depression detection approach," *computation,* vol. 13, no. 1, p. 9, 2025.
10. G. A. Pradnyana, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "Revealing depression through social media via adaptive gated cross-modal fusion augmented with insights from personality traits," *IEEE Access*, vol. 13, no. 7, pp. 133465–133482, 2025.
11. M. Hu, L. Xu, L. Liu, X. Wang, H. Li, and J. Yang, "Automatic depression detection network based on facial images and facial landmarks feature fusion," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal, 2024.
12. Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin, G. Huang, Z. Liang, L. Zhang, L. Li, H. Ding, Z. Zhang, and J. Wang, "Multimodal sensing for depression risk detection: Integrating audio, video, and text data," *Sensors*, vol. 24, no. 12, p. 3714, 2024.
13. C. Lu and X. Fu, "SentDep: Pioneering fusion-centric multimodal sentiment analysis for unprecedented performance and insights," *IEEE Access*, vol. 12, no. 2, pp. 21277–21286, 2024.
14. Y. Zhou, W. Han, X. Yao, J. Xue, Z. Li, and Y. Li, "Developing a machine learning model for detecting depression, anxiety, and apathy in older adults with mild cognitive impairment using speech and facial expressions: A cross-sectional observational study," *International Journal of Nursing Studies*, vol. 146, no. 10, p. 104562, 2023.
15. M. Milling, A. Baird, K. D. Bartl-Pokorny, S. Liu, A. M. Alcorn, J. Shen, T. Tavassoli, E. Ainger, E. Pellicano, M. Pantic, N. Cummins, and B. W. Schuller, "Evaluating the impact of voice activity detection on speech emotion recognition for autistic children," *Frontiers in Computer Science*, vol. 4, no. 2, p. 837269, 2022.
16. H. Wang, Y. Liu, X. Zhen, and X. Tu, "Depression speech recognition with a three-dimensional convolutional network," *Frontiers in Human Neuroscience*, vol. 15, no. 9, p. 713823, 2021.
17. R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine*, vol. 135, no. 8, p. 104499, 2021.
18. K. O. Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, and D. Ferreira, "Predicting Depression from Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study," *JMIR mHealth and uHealth*, vol. 9, no. 7, p. e26540, 2021.
19. S. Alghowinem, T. Gedeon, R. Goecke, J. F. Cohn, and G. Parker, "Interpretation of depression detection models via feature selection methods," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 133–152, 2023.