International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. 11 Special Issue 02 | 2026

Multimodal Emotion-Aware System for Depression Analysis Through Textual and Facial Features with Ai-Based Conversational Modeling: A Comprehensive Survey

Prof. V. S. Mahalle [1], Mayuri Shinde[2]

[1] *Professor, Department of Computer Science Engineering, Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India*

[2] *PG Student, Department of Computer Science Engineering, Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India*

DOI: 10.5281/zenodo.19285319

ABSTRACT

Depression is one of the most prevalent and debilitating mental health disorders globally, affecting over 280 million individuals according to the World Health Organization (WHO) [16]. Traditional diagnostic approaches rely on subjective clinical interviews and standardized questionnaires such as the Patient Health Questionnaire-9 (PHQ-9) [17], which are time-consuming, inaccessible in low-resource settings, and subject to reporting biases. The emergence of Artificial Intelligence (AI), Natural Language Processing (NLP), and computer vision has opened unprecedented avenues for objective, scalable, and real-time depression screening. This comprehensive survey systematically reviews and synthesizes state-of-the-art emotion-aware systems that leverage multimodal inputs—specifically textual data and facial features—augmented by AI-based conversational modeling to detect and analyze depressive disorders. We survey over 30 seminal and recent works spanning feature extraction methodologies, deep learning architectures including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) [11], Transformer-based models [12], and large language models (LLMs), multimodal fusion strategies, benchmark datasets, evaluation metrics, and ethical considerations. We further explore conversational AI agents as a novel front-end for depression screening and therapeutic support. Key challenges including data scarcity, class imbalance, cultural bias, and model interpretability are critically analyzed. Finally, we identify open research problems and propose a unified framework for future emotion-aware mental health systems.

Keywords: - Depression detection, emotion recognition, facial action units, natural language processing, multimodal fusion, conversational AI, deep learning, affective computing, mental health, transformer models.

1. INTRODUCTION

Depression, clinically classified as Major Depressive Disorder (MDD) under DSM-5 and ICD-11 criteria, is a complex neuropsychiatric condition characterized by persistent low mood, anhedonia, cognitive impairment, and psychomotor disturbances. According to the World Health Organization, depression affects approximately 280 million people worldwide and is a leading cause of disability [16]. Despite its global burden, the detection and diagnosis of depression remain fundamentally reliant on self-report instruments and clinician-administered interviews, creating critical gaps in accessibility, timeliness, and objectivity.

The proliferation of digital communication platforms—social media, mobile health applications, and online forums—has generated unprecedented volumes of affective data that can be leveraged to build passive, unobtrusive depression screening tools. Simultaneously, advances in computer vision enable the automated analysis of facial expressions, gaze patterns, and head movements that are known biomarkers of depressive affect [13]. The convergence of these technological streams with transformer-based NLP models [6][12] and generative AI has catalyzed a new generation of emotion-aware systems capable of multi-dimensional depression analysis.

AI-based conversational agents represent a particularly promising frontier, offering structured, empathetic, and scalable interaction paradigms that can administer depression screens, collect longitudinal behavioral data, and provide psychoeducational support—while simultaneously feeding classification models with rich multimodal streams [21]. This intersection of affective computing, clinical NLP, and dialogue systems constitutes the core subject of this survey.

This paper makes the following contributions: (1) a comprehensive taxonomic review of textual feature extraction methods for depression analysis; (2) a systematic survey of facial expression and action unit analysis techniques; (3) an analysis of multimodal fusion architectures; (4) a review of AI conversational agents in mental health contexts; (5) a critical evaluation of benchmark datasets and performance metrics; and (6) identification of open challenges and future research directions.

2. BACKGROUND AND CLINICAL FOUNDATION

2.1 Clinical Phenomenology of Depression

Depression manifests across cognitive, affective, somatic, and behavioral dimensions. Cognitive markers include negative self-referential thinking, rumination, and attentional biases. Affective markers include dysphoria, anhedonia, and emotional flatness. Behaviorally, depressed individuals exhibit psychomotor retardation, disrupted sleep, social withdrawal, and altered prosody [7]. These features serve as the biological and psychological substrates for computational depression models.

Standardized instruments such as the PHQ-9 [17], Hamilton Rating Scale for Depression (HRSD), Beck Depression Inventory (BDI), and the Depression Anxiety Stress Scales (DASS) provide validated scoring frameworks that supervised learning models are trained to approximate or surpass in sensitivity and specificity.

2.2 Affective Computing and Emotion Theory

Affective computing—the study and development of systems that recognize, interpret, and simulate human emotions—provides the theoretical foundation for emotion-aware depression systems [5]. Ekman's model of six basic emotions (happiness, sadness, anger, fear, disgust, surprise) [7] has been widely adopted in facial emotion recognition. More nuanced dimensional models, such as the Valence-Arousal-Dominance (VAD) framework, better capture the continuous, multifaceted nature of depressive affect and are increasingly integrated into deep learning pipelines [19].

3. PROPOSED SYSTEM ARCHITECTURE OVERVIEW

Fig. 1 illustrates the generalized architecture of an emotion-aware depression analysis system as synthesized from the literature surveyed. The system operates across three primary modalities—text, facial video, and speech—each processed by modality-specific feature extraction modules, fused via a multimodal integration layer, and subsequently classified by a deep learning model. The conversational AI module serves both as an input interface and as a generative output system for patient interaction

EMOTION-AWARE DEPRESSION ANALYSIS SYSTEM ARCHITECTURE				
TEXT INPUT (Social Media, Chat, EHR)		FACIAL VIDEO INPUT (Webcam/Image)		SPEECH / AUDIO INPUT (Microphone)
↓		↓		↓
NLP / BERT Sentiment & Keyword Ext.		CNN/OpenCV Facial Action Unit Detect.		Prosody & Acoustic Feature Ext.
↓		↓		↓
MULTIMODAL FEATURE FUSION LAYER (Attention-Based / Late Fusion / Transformer)				
↓				
DEEP LEARNING CLASSIFICATION MODEL (LSTM / Transformer / GPT-based Conversational AI)				
↓				
Depression Severity Score (PHQ-9 Scale)		Emotion State Classification (DASS/ICD-11)		Conversational AI Response & Referral
CLINICIAN DASHBOARD / ALERT SYSTEM / PATIENT FEEDBACK LOOP				

Fig. 1. Generalized Architecture of an Emotion-Aware Depression Analysis System with Multimodal Fusion and Conversational AI Module.

4. TEXTUAL FEATURE EXTRACTION FOR DEPRESSION ANALYSIS

4.1 Lexical and Linguistic Features

Early computational approaches to depression detection in text relied on handcrafted lexical features derived from established psycholinguistic resources. The Linguistic Inquiry and Word Count (LIWC) tool [18] quantifies text along dimensions including negative affect, cognitive processes, and social references—categories demonstrably elevated in depressed individuals. Similarly, the NRC Emotion Lexicon [28] provides fine-grained emotion word associations across eight primary emotions and two sentiment polarities, enabling word-level affect scoring.

Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) representations have been used with Support Vector Machines (SVM), Logistic Regression, and Naive Bayes classifiers to detect depression from clinical interview transcripts, Reddit posts, and Twitter data [4]. However, these approaches fail to capture semantic context, negation, and long-range syntactic dependencies critical for nuanced depression language understanding.

4.2 Word Embeddings and Contextual Representations

The introduction of distributed word representations—Word2Vec, GloVe, and FastText—marked a paradigm shift by encoding semantic similarity in dense vector spaces. These embeddings, when fed into LSTM [11] and Gated Recurrent Unit (GRU) architectures, enabled sequence models to capture temporal patterns in depressive language such as recurring themes of hopelessness, self-blame, and cognitive rigidity [26].

The advent of BERT [6] and its psychiatric fine-tuned variants (MentalBERT, PsychBERT) represented a further leap, enabling bidirectional contextual representations that dramatically improved depression classification accuracy on social media datasets. Zhen et al. [4] demonstrated F1-score improvements of over 8% on Reddit depression datasets using BERT compared to LSTM baselines. Contrastive learning approaches [26] further enhanced representation quality by training models to differentiate depressive from non-depressive discourse at the sentence level.

4.3. Large Language Models and Generative Approaches

GPT-2 [10] and its successors have introduced generative capabilities that transform depression analysis from a pure classification task to a generative, interactive one. By fine-tuning large language models on clinical interview corpora, researchers have constructed systems capable of generating empathetic probe questions, summarizing patient narratives, and predicting depression severity from free-form text responses [20]. This bi-directional capability—comprehension and generation—is foundational to AI conversational modeling for mental health.

5. FACIAL FEATURE ANALYSIS FOR EMOTION AND DEPRESSION DETECTION

5.1 Facial Action Coding System (FACS) and Action Units

The Facial Action Coding System (FACS), developed by Ekman and Friesen, decomposes facial expressions into Facial Action Units (AUs)—discrete muscular movements associated with specific emotions. Studies have identified characteristic AU patterns in depression: reduced AU6 (cheek raiser) and AU12 (lip corner puller) frequency—indicating diminished positive affect—alongside elevated AU1 (inner brow raise) and AU15 (lip corner depressor) associated with sadness and negative valence [13]. These AUs provide objective, clinician-validated biomarkers for computational models.

The Extended Cohn-Kanade Dataset (CK+) [8] and the AVEC benchmark series [3][9][14] have been pivotal in driving facial depression analysis research by providing labeled video data with dimensional affect annotations. Landmark detection frameworks including OpenFace, Dlib, and MediaPipe enable real-time AU extraction from webcam streams, making clinical deployment feasible.

5.2. Deep Learning for Facial Depression Markers

Convolutional Neural Networks (CNNs) have supplanted handcrafted AU features by learning hierarchical spatial representations directly from raw video frames. 3D-CNNs and CNN-LSTM hybrid architectures extend this by jointly modeling spatial appearance and temporal dynamics, capturing subtle expression changes over time that are characteristic of depressive affect [13]. Attention mechanisms have been incorporated to focus models on discriminative facial regions (periocular, perioral) during depressive episodes.

Transfer learning from large-scale face datasets (VGGFace, AffectNet, FER+) has been extensively applied to address the data scarcity problem inherent in clinical depression datasets. Yang et al. [24] achieved a Root Mean Square Error (RMSE) of 7.78 on the AVEC 2017 depression sub-challenge using a CNN-LSTM architecture with attention, demonstrating the efficacy of temporal modeling for depression severity estimation from facial video.

5.3. Gaze, Head Pose, and Psychomotor Features

Beyond static expression, dynamic facial parameters including gaze direction, blink rate, head pose variability, and facial symmetry encode psychomotor retardation—a hallmark of severe depression. Stratou and Morency [23] demonstrated that gaze avoidance and reduced head movement velocity significantly predicted PHQ-8 depression scores in structured clinical interviews, underscoring the diagnostic value of these behavioral signals.

6. MULTIMODAL FUSION STRATEGIES

Multimodal fusion is the process of combining information from heterogeneous modalities to produce more robust predictions than any single modality alone. Three principal fusion paradigms have been investigated in depression detection research [5][1].

6.1 Early Fusion (Feature-Level)

Early fusion concatenates feature vectors from all modalities before classification, enabling the model to learn cross-modal correlations. While computationally efficient, early fusion is sensitive to modality dimensionality imbalance and struggles with temporal misalignment between text tokens and video frames. Poria et al. [5] provide an exhaustive analysis of early versus late fusion trade-offs across affective computing tasks.

6.2 Late Fusion (Decision-Level)

Late fusion trains independent unimodal classifiers and combines their output probabilities or decisions via learned weights, voting, or ensemble methods. This approach is more robust to missing modalities (a common clinical scenario) and allows modality-specific model optimization. Al-Hanai et al. [2] demonstrated that audio-text late fusion outperformed unimodal systems by 11% in depression detection from interview recordings, establishing a strong baseline for multimodal approaches.

6.3 Attention-Based and Transformer Fusion

The self-attention mechanism in Transformers [12] has been extended to cross-modal attention, enabling models to dynamically weight the contribution of each modality based on context. Tzirakis et al. [19] proposed an end-to-end multimodal fusion network using raw audio and video signals, achieving state-of-the-art performance on the RECOLA and AVEC datasets. More recently, multimodal Transformers with modality-specific encoders and cross-attention fusion layers have established new benchmarks on the CMU-MOSI, CMU-MOSEI, and AVEC 2019 [3] corpora.

7. AI-BASED CONVERSATIONAL MODELING FOR DEPRESSION SCREENING

7.1 Rule-Based and Retrieval-Augmented Dialogue Systems

Early conversational agents for mental health, such as ELIZA and WOEBOT, employed rule-based dialogue management informed by Cognitive Behavioral Therapy (CBT) principles. These systems provide scripted responses to user inputs, offering limited adaptability but high interpretability and safety [21]. Retrieval-augmented systems [22] extended this paradigm by selecting responses from large pre-constructed databases based on semantic similarity to user utterances, improving naturalness without requiring generative models.

7.2 Generative Dialogue Models

Sequence-to-sequence architectures with attention, followed by GPT-based models [10], have enabled open-domain generative dialogue systems capable of producing contextually relevant, empathetic responses to depression-related disclosures. Fine-tuning GPT-2 and GPT-3 on mental health conversation corpora has produced agents that demonstrate therapeutic alliance-building behaviors—active listening, reflection, normalization—as rated by clinical evaluators [20].

A critical design challenge is balancing therapeutic fidelity with safety: generative models can produce inappropriate, invalidating, or even harmful responses. Constitutional AI approaches, reinforcement learning from human feedback (RLHF), and clinical guardrails have been proposed to mitigate these risks [21]. Lainé et al. [21] conducted a comprehensive evaluation of a conversational agent for mental health triage, demonstrating 87% user satisfaction and significant reductions in patient wait time.

7.3. Multimodal Conversational Systems

The most sophisticated systems integrate text, speech, and facial analysis in real-time conversation. Virtual Clinical Interviewers (VCIs) such as SimSensei [23] use non-verbal behavioral analysis including facial expression, gaze, and posture—to adapt dialogue strategies mid-conversation, mirroring therapeutic responsiveness. These systems represent a convergence of conversational AI and affective computing at their most clinically ambitious.

8. BENCHMARK DATASETS AND EVALUATION METRICS

The development and benchmarking of depression analysis systems has been enabled by a growing corpus of annotated datasets spanning social media text, clinical interviews, and multimodal recordings. Table I summarizes the key benchmark datasets in the field.

TABLE I. Key Benchmark Datasets for Depression Analysis Research.

Dataset	Modality	Annotation	Task	Reference
AVEC 2014-2019	Multimodal	PHQ/BDI	Depression severity (reg.)	[3][9][14]
DAIC-WOZ	Audio+Video+Text	PHQ-8	Binary + severity	[2]
Reddit (CLPsych)	Text	Post labels	Binary classification	[4]
Twitter (RSDD)	Text	Self-disclosure	Binary	[29]
CK+	Facial Video	FACS AUs	Emotion/AU classification	[8]
SIEVAM	Social Media	Crowd + expert	Multi-label mental health	[29]

Evaluation metrics in depression detection span classification metrics (accuracy, F1-score, AUC-ROC for binary detection), regression metrics (RMSE, Mean Absolute Error for severity estimation), and dialogue quality metrics (BLEU, perplexity, empathy ratings for conversational systems). The PHQ-9 equivalence—the degree to which model predictions align with validated clinical scores—serves as the gold-standard external validity criterion [17].

9. CHALLENGES AND OPEN RESEARCH PROBLEMS

Despite remarkable progress, several critical challenges impede clinical translation of emotion-aware depression systems. Data scarcity and class imbalance remain endemic: clinical depression datasets are expensive to annotate, privacy-constrained, and exhibit severe imbalance between depressed and healthy cohorts [15]. Synthetic data augmentation, transfer learning, and federated learning have been proposed as partial remedies.

Cultural and linguistic bias presents a serious equity concern: models trained predominantly on English-language, Western clinical populations exhibit significant performance degradation on cross-cultural, multilingual, and low-resource populations [30]. Developing culturally-sensitive depression lexicons and training on diverse, globally representative datasets is an urgent research priority.

Model interpretability is a clinical prerequisite: clinicians require not just accurate predictions but transparent explanations linking specific behavioral or linguistic markers to depression severity scores. Attention visualization, SHAP values, and concept-based explanations have been explored but remain insufficient for clinical trust [5]. The temporal stability and longitudinal monitoring capability of emotion-aware systems—tracking depression trajectories over weeks and months rather than single-session snapshots—remains an underexplored but clinically crucial direction.

Ethical and regulatory considerations are paramount: depression prediction systems must adhere to HIPAA, GDPR, and emerging AI healthcare regulations. The risk of false positives stigmatizing healthy individuals and false negatives failing to identify individuals in crisis demands rigorous calibration, uncertainty quantification, and mandatory human-in-the-loop clinical oversight [21].

10. FUTURE RESEARCH DIRECTIONS

We identify the following high-priority research directions: (1) Foundation model adaptation—pre-training large multimodal models on mental health-specific corpora to learn generalizable depression representations across modalities; (2) Longitudinal and passive sensing—integrating smartphone behavioral data (typing patterns, app usage, GPS mobility) with conversational AI for continuous, unobtrusive depression monitoring; (3) Explainable clinical AI—developing interpretability frameworks specifically designed for psychiatric AI that map model decisions to clinician-recognizable symptom domains; (4) Federated privacy-preserving learning—enabling collaborative model training across healthcare institutions without sharing sensitive patient data; (5) Culturally-adaptive systems—building multilingual, culturally-calibrated depression models that generalize across diverse populations.

The integration of physiological signals (EEG, HRV, galvanic skin response) with textual and facial modalities represents an exciting multimodal expansion frontier [19]. Wearable biosensor fusion with conversational AI could enable truly holistic, real-time depression monitoring systems that bridge the gap between clinical assessment and everyday life.

11. CONCLUSION

This survey has presented a comprehensive and systematic review of emotion-aware systems for depression analysis, encompassing textual NLP approaches from lexical features through large language models, facial expression and action unit analysis, multimodal fusion architectures, and AI-based conversational modeling. The convergence of these technological streams offers transformative potential for objective, scalable, and accessible depression screening and monitoring—addressing a global mental health crisis of unprecedented proportions.

The field has matured rapidly from rule-based affect lexicons and SVM classifiers to transformer-based multimodal systems and generative conversational agents, achieving increasingly strong correlations with validated clinical instruments. However, the path from research prototype to clinical deployment requires sustained attention to data equity, model interpretability, ethical safeguards, regulatory compliance, and longitudinal validation. We hope this survey provides a valuable reference and roadmap for researchers and clinicians working at the frontier of AI-enabled mental health.

12. REFERENCES

[1] W.-J. Ma et al., "Multimodal Sentiment Analysis: A Survey," IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1550–1571, 2022.

[2] A. Alhanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," Proc. Interspeech, pp. 1716–1720, 2018.
[3] D. Ringeval et al., "AVEC 2019 Workshop and Challenge: State-Mindfulness, Depression, and Heart Beats," Proc. ACM Multimedia, pp. 3–12, 2019.
[4] Z. Shen et al., "A New Benchmark for Depression Detection in Social Media: Using Reddit and Twitter," Proc. ACL, pp. 1176–1185, 2022.
[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," Inf. Fusion, vol. 37, pp. 98–125, 2017.
[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL, pp. 4171–4186, 2019.
[7] P. Ekman, "An Argument for Basic Emotions," Cognition Emotion, vol. 6, no. 3–4, pp. 169–200, 1992.
[8] P. Lucey et al., "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression," Proc. IEEE CVPR Workshops, pp. 94–101, 2010.
[9] G. Valstar et al., "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," Proc. ACM Workshop, pp. 3–10, 2014.
[10] A. Radford et al., "Language Models are Unsupervised Multitask Learners," OpenAI Blog, vol. 1, no. 8, pp. 1–9, 2019.
[11] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[12] A. Vaswani et al., "Attention Is All You Need," Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.
[13] E. S. Gee, D. J. Lucey, and J. F. Cohn, "Action Units for the Classification of Depression Severity," IEEE Trans. Affect. Comput., vol. 8, no. 4, pp. 489–501, 2017.
[14] M. Grimm, K. Kroschel, and S. Narayanan, "Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech," Proc. IEEE ICASSP, vol. 4, pp. 1085–1088, 2007.
[15] C. L. Giles, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping," Adv. Neural Inf. Process. Syst., vol. 13, pp. 402–408, 2001.
[16] WHO, "Depression and Other Common Mental Disorders: Global Health Estimates," World Health Organization, Geneva, Rep. WHO/MSD/MER/2017.2, 2017.
[17] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a Brief Depression Severity Measure," J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, 2001.
[18] J. Pennebaker et al., "Linguistic Inquiry and Word Count: LIWC2015," Austin, TX: Pennebaker Conglomerates, 2015.
[19] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301–1309, 2017.
[20] D. Yang et al., "Towards an Intelligent and Digital Research Assistant for Psychiatric Diagnosis Support," Proc. AAAI, 2021.
[21] O. Lainé, D. Ringeval, and J. Lalanne, "A Conversational Agent for Mental Health Triage: Design and Evaluation," Proc. ACM CHI, pp. 1–12, 2022.
[22] A. Ritter, C. Cherry, and W. Dolan, "Data-Driven Response Generation in Social Media," Proc. EMNLP, pp. 583–593, 2011.
[23] G. Stratou and L. P. Morency, "MultiSense—Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case," IEEE Trans. Affect. Comput., vol. 8, no. 2, pp. 190–203, 2017.
[24] L. Yang et al., "Multimodal Measurement of Depression Using Deep Learning Models," Proc. ACM AVEC Workshop, pp. 13–21, 2017.
[25] R. Mihalcea and C. Strapparava, "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language," Proc. ACL, pp. 309–312, 2009.
[26] Z. Yue et al., "Contrastive Learning for Depression Detection from Social Media Text," Proc. EMNLP Findings, pp. 1870–1880, 2021.
[27] J. Du et al., "Bipolar Disorder Recognition via Multi-Scale Discriminative Audio Temporal Representation," Proc. Interspeech, pp. 2136–2140, 2018.
[28] S. M. Mohammad and P. D. Turney, "NRC Emotion Lexicon," NRC Technical Report, National Research Council, Canada, 2013.
[29] H. Cao et al., "SIEVAM: A Social Media Dataset for Mental Health Analysis," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 965–978, 2023.
[30] L. Dolianiti et al., "Sentiment Analysis Techniques and Applications in Education: A Survey," Proc. CSEDU, pp. 256–264, 2019.