IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1947~1954
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1947-1954

Unimodal and multimodal techniques for depression diagnosis: a comprehensive survey

Swathy Jayasree[1], Yashawini Sridhar[2]
[1]Cambridge Institute of Technology, Visvesveraya Technological University, Belagavi, India
[2]Department of Computer Science and Engineering, Cambridge Institute of Technology, Bengaluru, India

Article Info

Article history:

Received Aug 15, 2025
Revised Jan 12, 2026
Accepted Jan 25, 2026

Keywords:

Deep learning
Depression
Electroencephalogram
Facial expression
Speech

ABSTRACT

Depression is a common and major mental health condition that affects individuals across all age groups and any backgrounds, severely reducing their physical, emotional, and cognitive functioning. It goes beyond typical mood swings and requires a timely and accurate diagnosis to prevent severe consequences such as suicidal tendencies, self-harm, and long-term mental decline. The improving performance of deep learning and machine learning techniques has significantly enhanced the speed and accuracy of depression diagnosis using both unimodal and multimodal features. This comprehensive study gives a complete overview of the unimodal and multimodal methods used to diagnose depression in its early stages. Additionally, this survey summarizes the dataset, methods, and limitations of previous work presented in the domain of depression diagnosis and serves as a suitable reference for future analysis.

Corresponding Author:

Swathy Jayasree
Cambridge Institute of Technology, Visvesveraya Technological University
Belagavi, India
Email: swathyj.cse@cambridge.edu.in

1. INTRODUCTION

Depression is a complex mental illness that impacts a person's behavior, emotions, thoughts, and physical health [1]. It can be defined by low mood, feeling sad, emotional emptiness, and loss of interest that can disturb daily routines and even lead to suicidal attempts and thoughts [2]. Symptoms of depression can include sadness, irritability, lack of enjoyment, loss of motivation, memory difficulty, and physical discomfort like headaches and retardation. Figure 1 illustrates the classification of depression symptoms.

Depression is classified into different types, including major depressive disorder (MDD), persistent depressive disorder (PDD), disruptive mood dysregulation disorder (DMDD), premenstrual dysphoric disorder (PMDD), seasonal affective disorder (SAD), prenatal depression, and postpartum depression [3]. Each type is defined by specific symptoms and duration; major depressive episodes continue for a minimum duration of two weeks. The causes of depression are complex, including the combination of genetics, prenatal stress and epigenetics, biological mechanisms, family dynamics, sociocultural influences, substance use, unhealthy behavior, and early-life trauma. A mother's maternal stress has been shown to influence fetal brain development, possibly highlighting the risk of mood disorders in later stages of life [4].

According to the World Health Organization (WHO), an estimated 300 million people globally suffer from depression, making it the world's major cause of disability [5]. Given the growing number of depression cases, there is an urgent need for effective and accurate diagnosis, especially in minor cases, to reduce suffering and speed up cost-effective treatment. Current diagnosis methods primarily depend on clinical interviews, such as semi-structured and fully structured interviews [6], which depend on the

professional expertise of doctors, as well as the active participation of patients. These limitations have led researchers to explore different approaches in this area. In recent years, machine learning and deep learning have significantly enhanced diagnostic capabilities, particularly in healthcare [7], [8]. Machine learning methods, such as support vector machine (SVM) and decision trees have a crucial role in prediction and classification tasks. However, deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can extract high-level features automatically from medical data, such as bio signals, magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET) scans, with less manual effort.

Studies have explored different unimodal and multimodal methods combined with machine learning and deep learning techniques to improve depression detection accuracy. Unimodal techniques consist of single modalities, including facial expression [9], speech [10], and electroencephalography (EEG) [11], and the multimodal approach [12]–[14] that combines multiple features to improve diagnostic accuracy. Figure 2 illustrates the different unimodal and multimodal methods used in this survey. In unimodal methods, the speech and EEG are integrated with deep learning utilizing CNNs for feature extraction and a long short-term memory network (LSTM) for analyzing the patterns. These methods capture both spatial and temporal features with improved feature representation. However, the primary limitation of unimodal methods is their inability to handle multiple features for diagnosing depression. These methods fail to utilize the complete capability of CNN, LSTM, and any other deep learning techniques.

In recent years, with the trend towards multimodality, an increasing number of studies have considered combining other modalities, especially EEG, speech, facial expressions, and eye tracking. Most of the multimodal approaches rely on dedicated feature extraction networks for each modality. This paper aims to present a comprehensive study, focusing on the advancements in machine learning and deep learning-based techniques for depression diagnosis through unimodal and multimodal methods.

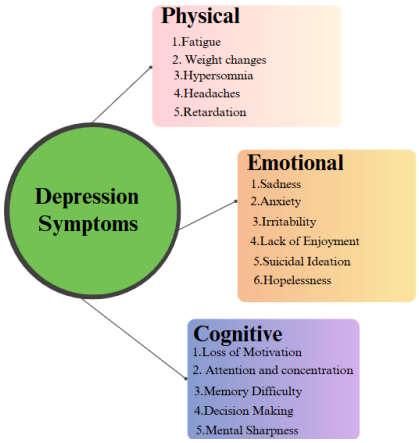

Figure 1. Classification of depression symptoms

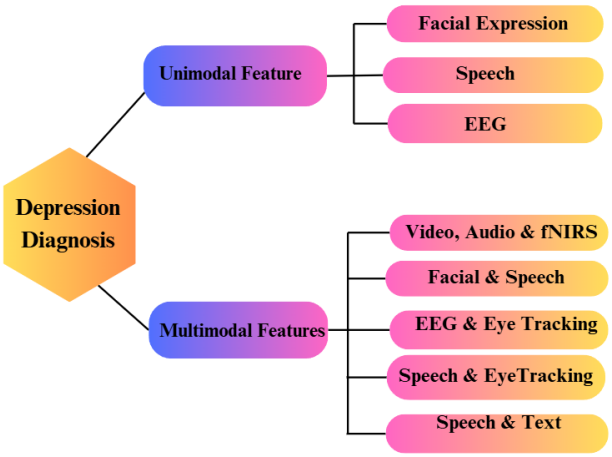

Figure 2. Different unimodal and multimodal features used in depression diagnosis

2. RELATED WORK

This section provides a detailed overview of unimodal and multimodal methods used for depression diagnosis using deep learning and machine learning techniques. This comprehensive analysis enables a good comparison of existing deep learning and machine learning approaches while revealing the gap and open research challenges that for further investigation. For each study, we discuss the features, datasets, methods, and research gap as follows.

2.1. Unimodal methods

Du *et al.* [15] introduce the multi-source chain depression recognition model is a deep learning technique designed to recognize depression using speech analysis. This model integrates LSTM networks with a one-dimensional CNN to extract features related to both speech production and perception. This model provides a deep understanding about verbal signs related to repression. In 2022, Rejaibi *et al.* [16] present a model that depends on Mel-frequency cepstral coefficient (MFCC) and a deep learning neural network model. This system used LSTM modules to derive advanced temporal features from speech, allowing for the

identification of depressive signals. Initially, the audio segments are processed with MFCCs to identify the frequency. This model presents a time-oriented approach that successfully fetches the speech fluctuations to depressive conditions. Kumar *et al.* [17] introduce a deep learning system that combines a modified visual geometry group (VGG)-16 model, LSTM, and fast Fourier transform (FFT) to identify stress in voice waveforms. FFT is used to capture changes related to stress by splitting the voice into its frequency components. LSTM gives a better contextual analysis by temporal features and Mel-filterbank representations. The VGG16 architecture is then used to classify these characteristics, improving the accuracy of differentiating stress-related speech patterns. Sardari *et al.* [18] present a model CNN-based autoencoder. This model automatically fetches important features from speech and reduces the manual effort. This model allows the autoencoder to retrieve depressive patterns from the speech.

Srimadhur and Lalitha [19] introduce a combined model using a CNN and an end-to-end CNN. CNN works with spectrograms to extract the visual representation of speech; the end-to-end CNN analyzes the speech input. For fetching features, both techniques use convolutional layers and max-pooling layers. This model gives a disadvantage, such as variations in speaker volume that impact spectrogram-based methods. Yin *et al.* [20] propose a speech-based model that is the integration of a transformer with a concurrent CNN. This model uses two techniques of low-level MFCC attributes, such as the transformer stream and the parallel CNN. The transformer stream uses linear attention to fetch the long-range temporal features and the parallel CNN fetches localized audible features. The output from these modules is passed through dense and SoftMax layers for classification. This integrated architecture permits the model to capture both global and local speech features associated with depression. Romero and Antolín [21] present an ensemble CNN framework for detecting depression automatically through speech analysis. The system uses log-spectrograms extracted from four-second audio clips to successfully capture vocal traits related to both frequency and temporal features. A series of CNN models is trained, and these outputs are combined using ensemble averaging methods. In 2021, Das and Naskar [22] propose a deep learning model that integrates MFCC feature extraction with spectrogram analysis, which is processed through a spectro CNN. The noise filtering and segmentation are used for pre-processing the input. The outputs from the MFCC and spectrogram are then fed into the neural network named do not disturb network (DNDNet) to perform the binary classification.

Zhu *et al.* [23] introduce a graph-oriented model using a graph convolutional network (GCN) called the graph input layer attention convolutional network. EEG recordings from the 128-channel HydroCel Geodesic Sensor Net (HCGSN). Important features such as spectral complexity and density were extracted from 105 electrodes. The final design consists of two GCN layers with leaky rectified linear unit (LeakyReLU) activation, batch normalization, and a dense output layer for final classification. Wang *et al.* [24] propose a multi-task approach; this model uses FFT, CNN, and transformer encoders to identify spectral and temporal characteristics. Li *et al.* [25] present a model for diagnosing depression using EEG; fuzzy labeling is used for feature selection. The brain functional connectivity, utilizing the phase lag index (PLI) gives the relationships between channels in EEG data. A project matrix is used to reduce the dimensionality of the feature, and a sparse regression model identifies the different features. A SVM is used for final classification. Shao *et al.* [26] propose a system based on a decentralized and centralized learning structure. This model consists of two main components: RegionalCalculationNet and GlobalCalculationNet. In the first module, EEG channels are classified based on brain regions, and spatial features are extracted using targeted attention and convolutional methods. These features are then given to the GlobalCalculationNet, which uses a multi-head attention mechanism to identify spatio-temporal relationships over the brain.

Ren and Song [27] present a graph-based model for collecting EEG recordings from 128-channels. This method involves converting EEG signals into brain networks at both individual and group levels by using different entropies along with graph-theoretical analysis. A dedicated deep learning model called physics-informed graph attention network (P-I-GAT), which is based on graph attention networks (GAT), integrates multi-rhythmic spatial attributes from these brain networks. This design allows for strong classification performance by capturing narrow spatiotemporal EEG patterns that depend on different levels of depression. Hou *et al*. [28] propose a deep learning architecture called the lightweight convolutional transformer neural network (LCTNN) aimed at recognizing depression through EEG data. The architecture consists of four different elements: the channel modulator, which uses Hjorth parameters to modify the weights of EEG channels; the temporal-spatial embedding, which gives detailed temporal and spatial details. The sparse attention, which amplifies computational efficiency, and the attention pool, which eliminates non-dominant features. Rafiei *et al.* [29] propose a deep learning model for the prediction of MDD using EEG signals. This model is based on the InceptionTime architecture, incorporating six inception modules, integrating bottleneck layers, and employing filter lengths of 10, 20, and 40 to capture both short-term and long-term features. To improve input features, a three-step EEG channel selection method, consisting of mean absolute difference (MAD), correlation coefficient analysis, and backward elimination, is implemented. Seal *et al.* [30] present a model to diagnose depression from EEG data. This model is the combination of five different convolutional

blocks that capture spatial-temporal features from different nodes. In the classification process, this model provided the different brain activity-related hemispheric asymmetries. Liu *et al.* [31] introduced a graph based method that combines the time frequency complexity with spatial topologies. This model divides the EEG signals into different frequencies, such as delta, theta, alpha, for fetching different entropy attributes. These features are preprocessed through a bidirectional long short-term memory (Bi-LSTM) network to understand about the temporal patterns. The brain graphs are created using Pearson correlation techniques. Both spatial and temporal representations are then input into a graph convolution network to diagnose the depression related signs. Zhang *et al.* [32] present a combined neural network model is the integration of 2D-CNN with LSTM networks. This model enables the system to fetch both spatial and temporal patterns from EEG signals.

Chung *et al.* [33] propose a model that uses a mobile application to accept the EEG signals from 57 individuals during 10-minute sessions. The EEG signal is divided into eight sub-bands, and each band is used to train a neural network based on LSTM. These techniques are integrated through a two-tier model, such as a combined network and multiple linear regression. The combined network that was trained on different combinations of bands and multiple linear regression is used to combine outputs for classification. Hu *et al.* [34] present a model that uses facial expressions from 62 individuals. Local face reconstruction technique is used to identify the facial regions of interest. These regions, the model derived main features such as local phase quantization on three orthogonal planes, action unit (AU) intensity, and measurement of head movements. These attributes are then input into a SVM classifier to differentiate depressive properties. Pan *et al.* [35] propose a deep learning model named the spatial-temporal attention (STA) depressive recognition network for identifying depression using facial expressions. This model uses the STA technique to create pixel-level attention vectors to get both facial and specific patterns. Li *et al.* [36] present a method with two features: first is the dual scale convolutional module (DSCM), to capture multi-scale facial features. This model implements 3×3 and 5×5 convolutions and the adaptive channel attention mechanism (ACAM), which highlights channel features using a learning pooling approach to improve selective capabilities. Lee and Park [37] propose a model using a region-based CNN. This system collects facial images with eye and lip regions via a chatbot interface. These images are processed to identify precise shifts in facial signals that are related to depression. Rajawat *et al.* [38] present a hybrid model that combines fuzzy logic with deep learning. This model uses a 3D-CNN to fetch spatio-temporal features from video frames. To avoid ambiguity present in facial signals, a fuzzy logic layer is combined into deep learning features. A brief overview of selected studies regarding depression diagnosis using unimodal methods is provided in Table 1.

2.2. Multimodal techniques

Wang *et al.* [39] introduce a multimodal feature perception and multiple cross-attention fusion, which is a hybrid model designed for the diagnosis of depressive episodes. The system analyzes facial video, audio, and functional near-infrared spectroscopy (fNIRS) data collected under consistent incentive tasks. For video input, a multi-scale CNN paired with a gated recurrent unit (GRU) is utilized to extract spatiotemporal behavioral characteristics; the audio data is transformed into Mel-spectrogram heatmaps and processed by a vision transformer to capture temporal-spectral prosody; and the fNIRS signals are interpreted using a multi-channel CNN to identify neurophysiological patterns. These features specific to each method are fused through a transformer-based cross-attention mechanism, which dynamically aligns and combines cross-modal dependencies for a semantically enhanced representation. Zhang *et al.* [40] present an innovative approach for identifying depression that combines hybrid fusion techniques with adaptive attention mechanisms. The multimodal encoding network architecture analyzes facial video frames, audio spectrograms, and MFCC features through three dedicated branches using CNNs and Bi-LSTMs to extract spatiotemporal patterns. The facial attributes capture micro-expressions and the temporal dynamics present in video frames, while the audio branches concentrate on speech pattern signals using MFCCs and emotional patterns derived from spectrograms. The output from each method contributes to a dynamic fusion via the attention decision fusion (ADF) module, which adaptively weighs the predictions specific to each model for effective integration. This hybrid method uses targeted feature representations without increasing the dataset size and aligns semantic dependencies through attention mechanisms.

Zhu *et al.* [41] propose a multimodal transformer network (MTNet) focused on detecting mild depression by combining EEG and eye-tracking information. This model employs statistical features, Hjorth parameters, and nonlinear descriptors derived from EEG signals, together with eye-tracking metrics such as fixation distribution, area of interest (AOI) sample percentages, and fixation duration. Features specific to these modalities are encoded and fused through a transformer-based structure, which captures dependencies both within and between the modalities. MTNet uses adaptive fusion while keeping the dataset size constant, and its design focuses on achieving semantic alignment between neurophysiological and behavioral signals for more accurate classification.

Uddin *et al.* [42] introduce a comprehensive multimodal network designed for estimating the severity of depression automatically, merging video-based facial movements with acoustic characteristics. The facial features are extracted using the Inception-ResNet-V2 model, while the innovative volume local directional structural pattern descriptor captures changes over time. For the perceptible component, the model uses a 1D residual CNN to identify local acoustic patterns from an encoder-decoder Bi-LSTM to analyze temporal speech characteristics. Both features are summarized through temporal attentive pooling, which gives important temporal segments and are combined using multimodal factorized bilinear pooling to smooth cross-modal interaction.

Jin *et al.* [43] propose a model that integrates facial and audio signals for depression diagnosis. The facial video component incorporates a spatio-temporal network that includes an attention mechanism to capture both global and local patterns. Attributes from the audio are obtained using MFCC, which are then converted into graphs and processed with a GCN in combination with LSTM.

Hemalatha *et al.* [44] introduce a model for diagnosing depression using speech and eye-tracking features. The eye tracking features give behavioral cues with spatial and temporal aspects. Speech features are captured using MFCC along with rhythmic markers such as pitch and duration to capture emotional variations. This system uses a multimodal CNN using an SVM classifier for analyzing speech features and XGBoost for determining eye tracking features.

Kumar *et al.* [45] present a multimodal method for diagnosing depression, including video, audio, and textual data. Spatial and temporal features are extracted using a multitask cascaded convolutional network and ResNet-18. Audio spectograms are generated with Librosa and encoded using RestNet-18. Each data is processed through dedicated networks such as CNN, RNN, transformer, and multi layer perceptron (MLP). Solieman and Pustozerov [46] present a deep learning method that uses both text and audio. The textual data is processed through natural language processing (NLP), then passed through CNN and LSTM to maintain semantic and sequential factors. A brief overview of selected studies regarding depression diagnosis using multimodal methods is provided in Table 2.

Table 1. Summary of depression diagnosis using the unimodal method

Sl. No	Features	Dataset	Method	Research gap
1.	Speech	DAIC-WOZ, MODMA [15]	CNN+LSTM	Overreliance on perception features
		DAIC-WOZ, RAVDESS, AVi-D [16]	LSTM+Transfer learning	Sensitive to noisy or informal speech
		SEMAINE [17]	FFT+LSTM+VGG-16	Combines all emotional states into "stress"
		DAIC-WOZ [18]	CNN-AE+SVM	Manual features still common: overfitting risk
		DAIC-WOZ, AVEC [19]	Spectrogram-based CNN and end-to-end CNN	Poor handling of severity levels
		DAIC-WOZ, MODMA [20]	Transformer+CNN	High computational cost
		DAIC-WOZ [21]	CNN+Ensemble averaging	Privacy and resource concerns
		DAIC-WOZ, RAVDESS, MODMA [22]	CNN+Dense network	Cross-cultural testing needed
2.	EEG	Private EEG [23]	GICN with functional connectivity	Long-range EEG connection is missed
		Mumtaz 2016, Arizona 2020 [24]	CNN FFT + transformer + Contrastive task	Focused only on MDD vs healthy
		MODMA [25]	Sparse regression+SVM	Relies on handcrafted features
		Private EEG (Lanzhou) [26]	RegionalNet+GlobalNet with attention	Fixed grouping may miss variability
		MODMA [27]	P-I-GAT for depression grading	Lacks multimodal clinical features
		MPHC, private EEG [28]	LCTNN with sparse attention modules	Fails to record EEG dynamics
		Mumtaz2017 [29]	InceptionTime with channel selection	No spatial EEG modeling
		PHQ-labeled EEG [30]	CNN+Pooling+Dense layers	Needs lightweight models for wearables
		MODMA [31]	Time-frequency+spatial graph fusion	Needs explainable AI for clinicians
		MODMA [32]	2D-CNN+LSTM+Dropout	Lacks interpretability for clinicians
		Single-channel (custom dataset) [33]	LSTM+2-tier MLR ensemble	Band-wise model complexity, explainability gaps
3.	Facial expression	Private dataset [34]	ROI face reconstruction+SVM	Relies on handcrafted features
		AVEC 2013/2014 [35]	STA mechanism+ResNet	Lacks multimodal inputs
		Custom dataset [36]	Dual-scale CNN+Channel attention	Unimodal limitations
		Private dataset [37]	Fast R-CNN	Single-frame analysis, no multimodal data
		Custom dataset [38]	3D-CNN+Fusion fuzzy layer	Static fuzzy rules, unimodal scope

Table 2. Summary of depression diagnosis using multimodal methods

Sl. No	Features	Dataset	Method	Research gap
1	Video, audio, and fNIRS	Custom dataset [39]	CNN + GRU, ViT, and Transformer fusion	Not considered the possible influence of individual differences in the neurophysiological reactions
2	Facial video frames and audio	AVEC2013 and AVEC2014 [40]	CNN + Bi-LSTM + Attention fusion	No neurophysiological signals
3	EEG and eye tracking	Custom dataset [41]	Transformer-based fusion	Lacks multimodal granularity
4	Audio and facial video	AVEC2013 and AVEC2014 [42]	InceptionResNet + 1D-CNN + Bi-LSTM	Lacks multimodal granularity
5	facial-video and audio	E-DAIC [43]	TSNet + GCN + LSTM	Binary only, lacks multimodal granularity
6	Speech and eye	Custom dataset [44]	CNN + SVM + XGBoost	Eye movement and unified temporal modeling across EEG, which limits its ability to capture comprehensive emotional dynamics in real-time situations
7	Speech and text	AVEC, DAIC, and E-DAIC [45]	CNN-RNN + transformer + voting	The dynamic temporal cross modality interactions are not explored
8	Speech and text	DAIC-WOZ [46]	CNN+LSTM	Lacks neuro-based features

3. CONCLUSION

This survey presents a complete analysis of recent unimodal and multimodal techniques used in depression detection, along with their methods, datasets, merits, and demerits. Advanced deep learning and machine learning models have played a crucial role in enhancing precision from depressive states with different features. Additionally, the current state-of-the-art publications throw light on multimodal techniques which are providing best results. Creating such a depression system with multimodal methods will improve the diagnosis systems accuracy and performance. This paper is expected to be a useful tool for researchers in depression.

FUNDING INFORMATION

No funding is raised for this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Swathy Jayasree	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Yashawini Sridhar	✓	✓				✓		✓	✓	✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Datasets utilized in this research are cited in reference [19], [25], [35].

REFERENCES

[1] A. Stringaris, "Editorial: What is depression?," *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 58, no. 12, pp. 1287–1289, 2017, doi: 10.1111/jcpp.12844.

[2] B. Levis *et al.*, "Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews," *British Journal of Psychiatry*, vol. 212, no. 6, pp. 377–385, 2018, doi: 10.1192/bjp.2018.54.
[3] M. Shehab *et al.*, "Machine learning in medical applications: a review of state-of-the-art methods," *Computers in Biology and Medicine*, vol. 145, 2022, doi: 10.1016/j.compbiomed.2022.105458.
[4] C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," *Multimedia Systems*, vol. 27, no. 4, pp. 599–613, 2021, doi: 10.1007/s00530-020-00694-1.
[5] X. Cao, L. Zhai, P. Zhai, F. Li, T. He, and L. He, "Deep learning-based depression recognition through facial expression: a systematic review," *Neurocomputing*, vol. 627, 2025, doi: 10.1016/j.neucom.2025.129605.
[6] A. Hassan and S. Bernadin, "A comprehensive analysis of speech depression recognition systems," in *Conference Proceedings - IEEE SOUTHEASTCON*, 2024, pp. 1509–1518, doi: 10.1109/SoutheastCon52093.2024.10500078.
[7] K. Elnaggar, M. M. El-Gayar, and M. Elmogy, "Depression detection and diagnosis based on electroencephalogram (EEG) analysis: a systematic review," *Diagnostics*, vol. 15, no. 2, 2025, doi: 10.3390/diagnostics15020210.
[8] U. Arioz, U. Smrke, N. Plohl, and I. Mlakar, "Scoping review on the multimodal classification of depression and experimental study on existing multimodal models," *Diagnostics*, vol. 12, no. 11, 2022, doi: 10.3390/diagnostics12112683.
[9] E. A. Stepanov *et al.*, "Depression severity estimation from multiple modalities," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services, Healthcom 2018*, 2018, pp. 1–6, doi: 10.1109/HealthCom.2018.8531119.
[10] M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnykova, "Multimodal data fusion for depression detection approach," *Computation*, vol. 13, no. 1, 2025, doi: 10.3390/computation13010009.
[11] J. E. R. Bernard, "Depression: a review of its definition," *MOJ Addiction Medicine & Therapy*, vol. 5, no. 1, pp. 6–7, 2018, doi: 10.15406/mojamt.2018.05.00082.
[12] L. Orsolini *et al.*, "Understanding the complex of suicide in depression: from research to clinics," *Psychiatry Investigation*, vol. 17, no. 3, pp. 207–221, 2020, doi: 10.30773/pi.2019.0171.
[13] F. Benazzi, "Various forms of depression," *Dialogues in Clinical Neuroscience*, vol. 8, no. 2, pp. 151–161, 2006, doi: 10.31887/dcns.2006.8.2/fbenazzi.
[14] M. Bembnowska and J. J.-Ochojska, "What causes depression in adults?," *Polish Journal of Public Health*, vol. 125, no. 2, pp. 116–120, 2015, doi: 10.1515/pjph-2015-0037.
[15] M. Du *et al.*, "Depression recognition using a proposed speech chain model fusing speech production and perception features," *Journal of Affective Disorders*, vol. 323, pp. 299–308, 2023, doi: 10.1016/j.jad.2022.11.060.
[16] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, 2022, doi: 10.1016/j.bspc.2021.103107.
[17] A. Kumar, M. A. Shaun, and B. K. Chaurasia, "Identification of psychological stress from speech signal using deep learning algorithm," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 9, 2024, doi: 10.1016/j.prime.2024.100707.
[18] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using convolutional autoencoder," *Expert Systems with Applications*, vol. 189, 2022, doi: 10.1016/j.eswa.2021.116076.
[19] N. S. Srimadhur and S. Lalitha, "An end-to-end model for detection and assessment of depression levels using speech," *Procedia Computer Science*, vol. 171, pp. 12–21, 2020, doi: 10.1016/j.procs.2020.04.003.
[20] F. Yin, J. Du, X. Xu, and L. Zhao, "Depression detection in speech using transformer and parallel convolutional neural networks," *Electronics*, vol. 12, no. 2, 2023, doi: 10.3390/electronics12020328.
[21] A. V.-Romero and A. G.-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, 2020, doi: 10.3390/e22060688.
[22] A. K. Das and R. Naskar, "A deep learning model for depression detection based on MFCC and CNN generated spectrogram features," *Biomedical Signal Processing and Control*, vol. 90, 2024, doi: 10.1016/j.bspc.2023.105898.
[23] J. Zhu *et al.*, "EEG based depression recognition using improved graph convolutional neural network," *Computers in Biology and Medicine*, vol. 148, 2022, doi: 10.1016/j.compbiomed.2022.105815.
[24] Y. Wang, S. Zhao, H. Jiang, S. Li, T. Li, and G. Pan, "M-MDD: a multi-task deep learning framework for major depressive disorder diagnosis using EEG," *Neurocomputing*, vol. 636, 2025, doi: 10.1016/j.neucom.2025.130008.
[25] Y. Li, Y. Fang, X. Ren, and L. Gao, "EEG-based depression recognition using feature selection method with fuzzy label," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 3, 2024, doi: 10.1016/j.jksuci.2024.102004.
[26] X. Shao, M. Ying, J. Zhu, X. Li, and B. Hu, "Achieving EEG-based depression recognition using decentralized-centralized structure," *Biomedical Signal Processing and Control*, vol. 95, 2024, doi: 10.1016/j.bspc.2024.106402.
[27] S. Ren and J. Song, "A graph-based method for automatic graded diagnosis of depression using EEG signals," *Biomedical Signal Processing and Control*, vol. 100, 2025, doi: 10.1016/j.bspc.2024.106973.
[28] P. Hou, X. Li, J. Zhu, and B. Hu, "A lightweight convolutional transformer neural network for EEG-based depression recognition," *Biomedical Signal Processing and Control*, vol. 100, 2025, doi: 10.1016/j.bspc.2024.107112.
[29] A. Rafiei, R. Zahedifar, C. Sitaula, and F. Marzbanrad, "Automated detection of major depressive disorder with EEG signals: a time series classification using deep learning," *IEEE Access*, vol. 10, pp. 73804–73817, 2022, doi: 10.1109/ACCESS.2022.3190502.
[30] A. Seal, R. Bajpai, J. Agnihotri, A. Yazidi, E. H.-Viedma, and O. Krejcar, "DeprNet: a deep convolution neural network framework for detecting depression using EEG," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021, doi: 10.1109/TIM.2021.3053999.
[31] W. Liu, K. Jia, and Z. Wang, "Graph-based EEG approach for depression prediction: integrating time-frequency complexity and spatial topology," *Frontiers in Neuroscience*, vol. 18, 2024, doi: 10.3389/fnins.2024.1367212.
[32] J. Zhang, B. Xu, and H. Yin, "Depression screening using hybrid neural network," *Multimedia Tools and Applications*, vol. 82, no. 17, pp. 26955–26970, 2023, doi: 10.1007/s11042-023-14860-w.
[33] K. H. Chung, Y. S. Chang, W. T. Yen, L. Lin, and S. Abimannan, "Depression assessment using integrated multi-featured EEG bands deep neural network models: leveraging ensemble learning techniques," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1450–1468, 2024, doi: 10.1016/j.csbj.2024.03.022.
[34] B. Hu, Y. Tao, and M. Yang, "Detecting depression based on facial cues elicited by emotional stimuli in video," *Computers in Biology and Medicine*, vol. 165, 2023, doi: 10.1016/j.compbiomed.2023.107457.
[35] Y. Pan *et al.*, "Spatial–temporal attention network for depression recognition from facial videos[Formula presented]," *Expert Systems with Applications*, vol. 237, 2024, doi: 10.1016/j.eswa.2023.121410.
[36] M. Li, Y. Wang, C. Yang, Z. Lu, and J. Chen, "Automatic diagnosis of depression based on facial expression information and deep convolutional neural network," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 5, pp. 5728–5739, 2024, doi: 10.1109/TCSS.2024.3393247.

[37] Y. S. Lee and W. H. Park, "Diagnosis of depressive disorder model on facial expression based on fast R-CNN," *Diagnostics*, vol. 12, no. 2, 2022, doi: 10.3390/diagnostics12020317.
[38] A. S. Rajawat, P. Bedi, S. B. Goyal, P. Bhaladhare, A. Aggarwal, and R. S. Singhal, "Fusion fuzzy logic and deep learning for depression detection using facial expressions," *Procedia Computer Science*, vol. 218, pp. 2795–2805, 2022, doi: 10.1016/j.procs.2023.01.251.
[39] Y. Wang, T. Qu, W. Zhu, Q. Wang, Y. Cao, and R. Gui, "A hybrid model using multimodal feature perception and multiple cross-attention fusion for depressive episodes detection," *Information Fusion*, vol. 124, 2025, doi: 10.1016/j.inffus.2025.103354.
[40] X. Zhang, B. Li, and G. Qi, "A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method," *Biomedical Signal Processing and Control*, vol. 96, 2024, doi: 10.1016/j.bspc.2024.106552.
[41] F. Zhu, J. Zhang, R. Dang, B. Hu, and Q. Wang, "MTNet: multimodal transformer network for mild depression detection through fusion of EEG and eye tracking," *Biomedical Signal Processing and Control*, vol. 100, 2025, doi: 10.1016/j.bspc.2024.106996.
[42] M. A. Uddin, J. B. Joolee, and K. A. Sohn, "Deep multi-modal network based automated depression severity estimation," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2153–2167, 2023, doi: 10.1109/TAFFC.2022.3179478.
[43] N. Jin, R. Ye, and P. Li, "Diagnosis of depression based on facial multimodal data," *Frontiers in Psychiatry*, vol. 16, 2025, doi: 10.3389/fpsyt.2025.1508772.
[44] S. Hemalatha, K. Jothimani, K. Swathi, S. Shibinta, W. J. Selvakumar, and D. Sathish, "Multimodal approach for depression detection: integrating speech and eye ball movement data," in *IEEE 2024 1st International Conference on Advances in Computing, Communication and Networking, ICAC2N 2024*, 2024, pp. 1011–1019, doi: 10.1109/ICAC2N63387.2024.10894891.
[45] P. Kumar, S. Misra, Z. Shao, B. Zhu, B. Raman, and X. Li, "Multimodal interpretable depression analysis using visual, physiological, audio and textual data," in *2025 IEEE Winter Conference on Applications of Computer Vision, WACV 2025*, 2025, pp. 5305–5315, doi: 10.1109/WACV61041.2025.00518.
[46] H. Solieman and E. A. Pustozerov, "The detection of depression using multimodal models based on text and voice quality features," in *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2021*, 2021, pp. 1843–1848, doi: 10.1109/ElConRus51938.2021.9396540.

BIOGRAPHIES OF AUTHORS

Swathy Jayasree earned her Bachelor of Technology (B.Tech. degree) in Computer Science and Engineering from Kerala University, in 2013. She has obtained her master's degree in M.Tech. (Computer and Information Science) from Cochin University, Kerala in 2016. Currently she is a research scholar at Visvesvaraya Technological University, Belagavi doing her Ph.D. in Department of Computer Science and Engineering, Cambridge Institute of Technology, KR Puram-Research Resource Centre and also working as assistant professor in Cambridge Institute of Technology, KR Puram. She has attended many workshops and induction programs conducted by various universities. Her areas of interest are machine learning, deep learning, artificial intelligence, and image processing. She can be contacted at email: swathyj.cse@cambridge.edu.in.

Yashawini Sridhar is an associate professor in the Department of Computer Science and Engineering at Cambridge Institute of Technology, Bangalore with an experience of 15 years in teaching. She is qualified in bachelor of engineering (B.E. degree) in Information Science and Engineering from Visvesvaraya Technological University, Belagavi and Master degrees in Software Engineering, from Visvesvaraya Technological University, Belagavi, and Ph.D. specialization in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi. Her areas of interest are NLP, AIML, deep learning, computer vision, and image processing. She can be contacted at email: yashaswini.cse@cambridge.edu.in.