Generative AI–Based Multilingual Multimodal Framework for Depression Detection

Uswa Ashraf, Hamid Ghous, Mubasher H. Malik, Majid Khawar

Department of Computer Science, University of Southern Punjab, Pakistan

Abstract: Depression is a prevalent psychological condition that is not easy to detect at an early stage due to its multipolar nature in terms of clinical manifestations and subjective clinical diagnoses. Recent advancements in generative artificial intelligence models, deep learning, and machine learning techniques have made it feasible to digitally identify depression based on how the illness manifests itself in speech, facial expressions, text, and physiological and behavioral characteristics. The study examines and discusses the use of large language, multimodal, and unimodal models for digital depression identification in various multilingual contexts. The technology consistently outperforms unimodal systems, according to the results, with enhanced transformer and cross-attention architecture performance in cross-modal relationship capture, a crucial component of clinical decision support. Large language models have been shown to have potential applications in few-shot learning, multilingual analysis, transcript-based severity estimation, data generation for simulation, and transparent clinical decision support systems. The current review aims to provide a systematic overview of the current methodologies and identify the areas of future research; however, much work needs to be done to have such applications extensive in number, culture-independent, and reflecting the ordinal level severity.

Keywords: Depression Detection, Artificial Intelligence (AI), Generative AI, Multimodal Models, Multilingual Analysis

Email: mubasher@usp.edu.pk

1. Introduction

Depression is a severe mental disorder that affects emotional, cognitive, and behavioral functions, and remains one of the formidable challenges for early detection in clinical practice. To overcome the subjectivity of traditional approaches, a wide variety of machine learning techniques have been explored for automated depression detection. Early studies using ML were focused on speech analysis and revealed that acoustic features like pitch, pauses, articulation, and speaking rate are closely associated with depressive symptoms, as presented in the analysis of clinical interview-based speech [1] and higher-order spectral speech analysis on the DAIC-WOZ dataset. Traditional classifiers such as SVM, AdaBoost, and CNN have been successfully applied in the speech signal for depression classification and severity prediction [2][3]. Concurrently, text-based methods were utilized to identify linguistic and affect patterns in social and clinical texts, in which CNN and RNN learning models successfully predicted depression from short,

unstructured text of [4] and Arabic social media analyses emphasized regional, depressive expression [5].

Moreover, visually-oriented approaches analyzed facial expression and facial action units in video, demonstrating substantial differences in sadness and happiness expression in depressed versus non-depressed participants [6], in addition to improved depression severity assessment via facial image sequences by deep learning techniques, such as DepNet [7]. Although unimodal methods gave some insight, they frequently yielded results that did not encompass every aspect of depression. Therefore, current studies are now more inclined to more general deep learning models that use audio, image, and text information together. The use of cross-attention models, along with transformer-based fusion models, led to even greater performance improvements in clinical depression detection tasks compared to unimodal models [8], although more improved methods were brought about by cross-attention models with transformer-based fusion separators [9] [10]. Given that large language models (LLMs) have become a trend, methods in depression detection have improved to include multi-language, few-shot, and explainable models, where LLM methods in speech detection or transcription gave high F1 scores in Chinese, Italian, and French cultures independently without any training[11]. LLMs have been applied effectively in depression severity estimation from transcripts of clinical interviews [12] and in improving model performance by generating synthetic data [13]. This review paper aims to give a comprehensive analysis of current methods available for depression detection using ML (machine learning), multimodal, and large language models.

- To survey the state-of-the-art of ML (machine learning) approaches, DL (deep learning) approaches, multimodal approaches, and approaches based on the use of LLM
- To effectively assess the performance of unimodal and multimodal strategies utilizing speech, text, vision, and physiological signals.
- That is, to summarize key datasets, feature extraction, as well as fusion techniques that have previously been used in various studies.
- For analyzing the use of large language models within the context of multilingual few-shot and explainable depression recognition.

For determining major challenges, the identification of gaps in research, such as data causalities, data biases, or data practicality. Providing an aid for comparison in future research for clinical

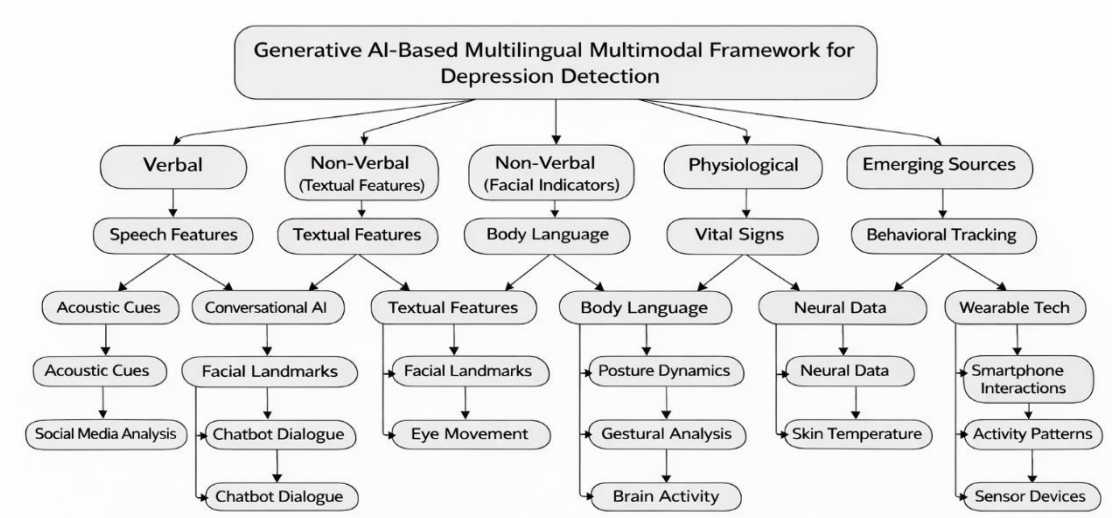

Figure 1. Generative AI-Based Multilingual Multimodal Framework for Depression Detection.

This figure illustrates the proposed framework integrating multiple data modalities for depression detection, including verbal cues (speech and conversational AI), non-verbal textual and facial indicators (textual features, facial landmarks, eye movement, and body language), physiological signals (vital signs and neural data), and emerging data sources (behavioral tracking through wearable devices and smartphone interactions). The framework highlights how diverse multimodal inputs can be jointly leveraged using generative AI models to enable more accurate, robust and scalable depression

use.

2. Literature Review

The taxonomy of the literature reviewed is done in a hierarchical and systematic manner to clearly highlight the research trends in the detection of depression. At the top level, the literature is categorized based on the methodological approach, namely DL (Deep Learning), ML (Machine Learning), and LLMs (Large Language Models), which represents the evolution from traditional learning approaches to advanced generative models. categorized based on the methodological approach, namely DL (Deep Learning), ML (Machine Learning), and LLMs (Large Language Models), which represents the evolution from traditional learning approaches to advanced generative models. These categories are further divided based on the type of dataset used, which includes text-only data, interview-based data, and multimodal data that combine textual, acoustic, visual, or physiological signals. further divided based on the type of dataset used, which includes text-only data, interview-based data, and multimodal data that combine textual, acoustic, visual, or physiological signals. This taxonomy assists in understanding the way in which various computational approaches utilize different data sources and also provides a

systematic outlook on the strengths, weaknesses, and research emphasis of existing studies in this area.

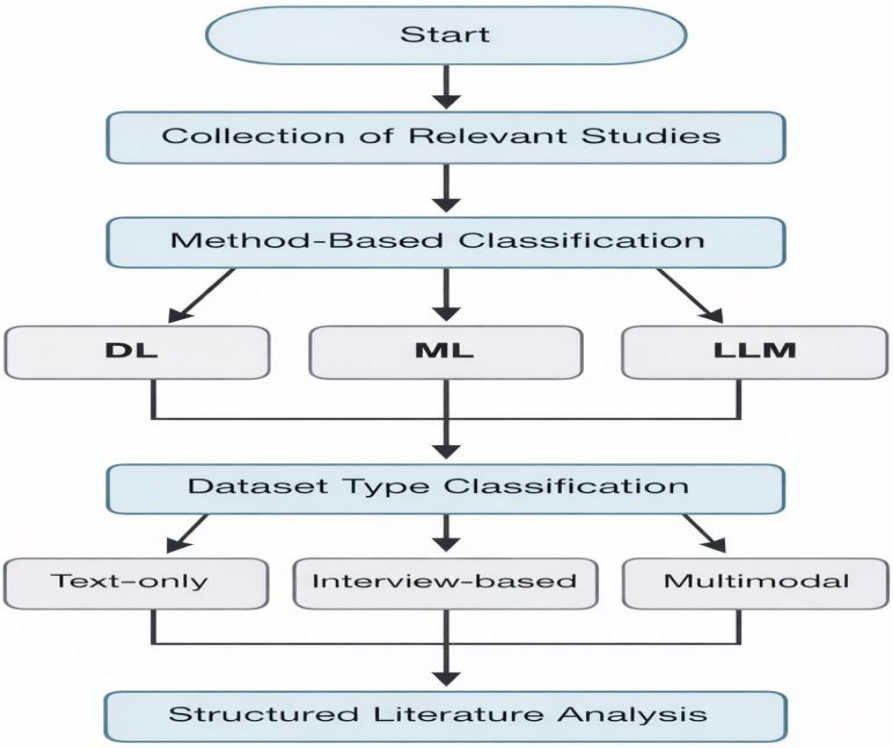

Figure 2. Flow diagram illustrating the process of literature classification based on methods and dataset modalities.

Minor resource language to detect depression and ignore the distinction between read speech and spontaneous speech. The deep learning techniques were used to combine the audio feature and textual embedding to enhance diagnostic accuracy. The hybrid model and the multimodal approach were employed, which integrates LSTM (Long Short-Term Memory) and a convolutional neural network to analyze audio in detail, and a pre-trained BERT model to provide textual clues. Based on a corpus of 288 recordings and 64 cases of depression, which were clinically diagnosed [14]. The Methodology has an accuracy of 94.3% and an F-score of 94.51% that is better in the detection of depression in both read and spontaneous speech. Deep learning algorithms were used for detecting depression, which combined the textual and audio features of patient responses. The DAIC-WOZ dataset was used, and then the audio and text features from the conversation were extracted. Method consist of three components: textual CNN, audioCNN, and a hybrid model with LSTM/BI-LSTM was applied. The results showed

that audioCNN detected the depression with an accuracy of 98%, textCNN 92%, and BI-LSTM showed better learning rate with an accuracy of 88% as compared to others [8]. A comparative evaluation of deep learning models for depression detection. The dataset used for the experiment was CLPsych2015 and the Bell Let's Talk Twitter dataset. To improve the feature learning word embedding, including skip-gram, CBow, and trainable random were evaluated. Four neural network models, three CNN (CNNwithmax, MultichannelCNN, multichannelpoolingCNN), and one RNN were used to evaluate the performance of depression detection. The results showed that CNN-based models were outperformed with an accuracy of 87.95% [4]. Depression Detection by combining facial, speech, and textual signals. Data collected during GPT-2 powered Chatbot interviews as audio, video, and text. Features from different modalities were fused using a multi-head cross-attention network to classify depression. Results showed in all internal scenarios Area under Curve >0.95 and accuracy >0.93, but in Chatbot interviews, AUC 0.999 outperformed all unimodal and bimodal [9].

Speech signals recorded via smartphones for depression detection. Audio samples from 318 participant with depression and healthy control to analyzed to determine their association with depression using a smartphone. The method used three main approaches: Conventional ML on acoustic features, proposed CNN using log Mel spectrogram, and pretrained deep models. The results showed that the proposed CNN achieved 78.14% accuracy, outperforming conventional ML and pretrained models [15]. Depression was predicted from the facial image sequence from videos by using a deep learning-based approach. Experiments were carried out by using the two databases AVEC2013 and AVEC2014, and 150 clips from each dataset were used. BDI-II is used as labelling for each sample. The DepNet model was used on the dataset to determine depression severity in a video sequence with fewer facial images. A feature aggregation model to aggregate low-level features of video data. The result showed the RMSE between predicted values and the BDI-II scores is 9.17 and 9.01 on the datasets. So the model performed better than the single-image model and state-of-the-art methods [7]. Articulatory coordination features (ACFS) were developed to capture changes of neuromotor coordination that happen as a result of psychomotor slowing, a necessary feature of Major Depressive Disorder. The dataset used for the experiment consists of 472 (35 speakers) and 753 (105 speakers) FS recordings from MD-1 and MD-2, respectively. CNN was an openSMILE feature to serve as a baseline model for a session-wise classification, and RNN to derive session-level predictions based on segment- level

predictions. The result showed that the RNN model trained using ACFS achieved a relative improvement of 27.47% in unweighted Average Recall (UAR) as compared to the others [16]. Lightweight cross-modality multimodal fusion model to detect depression on text and audio data. The datasets are DAIC-WOZ (English), EATD-Corpus (Chinese), and the Korean depression dataset. MLP Mixer is used to train audio features, and the transformer model is used to train textual features. The model scored an F1 Score of 0.67, 0.81, and 0.61 on the English, Chinese, and Korean datasets. The cross-language experiments prove that the proposed model can be used in other languages, even when the model is not trained. These Findings suggest that the model can combine nonlinguistic audio features and linguistic textual features using a multimodal strategy[17]. An end-to-end AI framework that automatically extracts depression cues from diary recordings. The data set used for the detection of depression includes 28 video recordings from the Symptoms Media dataset and 27 recordings from the DAIC-WOZ dataset. The author compared the presence of the extracted features between recordings of individuals with and without depressive disorders. The framework was successfully differentiated between individuals with or without depression [18]. A lightweight multimodal method for detecting depression severity based on speech and video signals. The Extended Distress Analysis interview corpus (E-DAIC) was used with 275 samples labeled with PHQ-8 scores. The model, an LSTM-based multimodal network with cross-model fusion, was used. The model achieved 83.86% accuracy with a very small model size of 0.52mb. Future work includes improving feature learning, handling gender imbalance, and clinical validation with hospitals [19].

To improve the precision of depression diagnosis using representation learning and knowledge transfer methods. DAIC-WOZ datasets were tested, which include the various data modalities of textual, imaging, and audio. The model used framework RLTK-MDD (representation learning and knowledge transfer for multimodal depression diagnosis), which understands the meaningful representation from multimodal data and shows the effectiveness of knowledge transfer from pre-trained language models. This enhances the performance of the diagnosis. Findings revealed that RLKT-MDD was more effective than the traditional and convolutional diagnostic approach and enhanced the accuracy on the DAIC-WOZ dataset [20]. To examine the nonverbal behavior of depressed people. The datasets are acoustic and visual features of 961 vlogs of approximately 160 hours with 816 different speakers. A transformer-based multimodal deep learning model that uses a cross-attention mechanism to generate a multimodal representation to detect depression.

Findings indicated that the model trained on D-vlog attains better depression detection performance compared to the DAIC-WOZ on both datasets [10]. An integrated multimodal transfer model that combines EEG signals and interview data to improve depression detection. Data were dataset mental disorder analysis dataset (MODMA) and the DAIC-WOZ. The audio information adds linguistic and paralinguistic information. Using multi-head cross attention, the framework can capture intra- and inter-modal correlation to improve depression detection. Findings indicated that the model had a 4.7 percent accuracy and 10 percent precision improvement on the MODMA and DAIC-WOZ datasets compared to the art method [21]. Another framework named IIFDD to detect depression based on multi-type data gathered by IoMT devices. The model does not use a single type of feature, but rather integrates low-level handcrafted features with high-level deep features in audio, text, and video. The model initially acquires the semantic content within each modality and then applies an inter-modal fusion module with attention mechanisms to integrate information between modalities to assist the system in comprehending speech, facial expression, and text emotional cues. The method was tested on two Chinese datasets of depression and achieved state-of-the-art performance [22]. AVTF-TBN model, which examines facial expressions, voice features, and spoken text at the same time. A special setup with reading and interview tasks was designed for naturally eliciting the emotional responses of participants. The system was tested on 1911 individuals, and it was shown that using all three modalities is far more effective than using only one. The results showed that the model exhibited the best scores when using a combination of both tasks with a high F1-score, precision, and recall. On the whole, the method demonstrates that multimodal sensor data can significantly help to detect the risk of depression early [23].

Multimodal used EEG-MRI data to develop a diagnostic framework for assessing mental disorders. The dataset used clinical EEG time series and MRI neuroimaging data collected from patients with various mental disorders. Model used Hierarchically Aligned Dual space transformer (HADST) for EEG-MRI fusion. Contextually Regularized Revers Consistency Training (CRRCT) mitigates domain shift and noise-related challenges using a bidirectional consistency paradigm. The results showed that mental effectively captures spatiotemporal biomarkers across modalities, improving diagnostic accuracy and generalization metrics [24]. How a person's speech changes might indicate how severe their depression is. Dataset used as a recorded talk with depressed people over several weeks, and looked at traits including pitch and

speech pauses. They found that patients with more severe depression paused more frequently and spoke more slowly. Interestingly, the interviewer's voices changed depending on how depressed each person was. Using these voice traits, the computer was able to determine the degree of depression with accuracy. The results showed that voice-based analysis could be a simple and discrete way to track depression[1]. A multimodal framework to enhance the detection of depression from clinical speech interview data. The dataset used three real-world clinical interview contained audio, visual, and dialogue. The model used a multimodal dialogue level transformer with sequential positional encoding, question context vectors, and adversarial learning using a gradient reversal layer. The result showed that the model outperformed bias-sensitive approached across the dataset [25].

References	Dataset	Method	Results	Limitation/Future work
[14]	228 recordings (64 depressed read + spontaneous speech)	Audio: LSTM+CNN, Text: BERT, decision-level fusion	94.30% accuracy; spontaneous speech is more discriminative	Small dataset; extend to diverse languages and real-world data
[8]	**DAIC-Woz** depression interview dataset	Textual CNN, Audio CNN, Hybrid LSTM & Bi-LSTM	Audio CNN 98% accuracy; Text CNN 92%; Bi-LSTM better learning	Limited metrics and dataset size; improve generalization
[4]	Twitter posts of users labeled for depression	Deep neural networks with multiple NLP architectures	Accuracy ranges from 87% to 94% across CNN, BiLSTM, and hybrid models, with the best models 94% accuracy on short tweets	Data scarcity and annotation limitations
[9]	Internal: 270 samples; External: 100 samples (audio-video-text)	Multimodal deep learning with cross-attention + Chatbot interviews	Internal AUC > 0.95, Chatbot AUC 0.999, external AUC 0.978	No longitudinal data, reduced performance in externally test on severe cases is needed
[15]	153 MDD patients + 165 healthy controls	CNN on log-Mel spectrograms vs. traditional ML	Best accuracy: 78.14% using the proposed CNN	Single language & read speech only

[7]	AVEC2013, AVEC2014	**DepNet**(Pre-trained CNNs+Feature Aggregation)	RMSE 9.17 (AVEC2013), 9.01 (AVEC2014)	Facial-only modality; limited datasets
[16]	Speech recordings from clinical-interview sessions(FS recording from MD-1 and MD-2)	Articulatory Coordination Features (ACFs) **CNN** for segment-level classification, **LSTM (RNN)** for session-level severity prediction.	Achieves 27.47% relative improvement in (UAR) at the session level	Limited datasets; real-time validation needed
[17]	DAIC, EATD, Korean depression datasets (English/Chinese/Korean)	Lightweight cross-modality multimodal fusion (audio + text)	F1: 0.67 (EN), 0.81 (CN), 0.61 (KR)	Performance varies by dataset
[18]	Symptom Media (28) + DAIC-WOZ (27) videos	End-to-end multimodal feature extraction	Accuracy =83%–88% AUC > 0.85 across modalities	Small dataset needs larger, real-world validation
[19]	Audio and video recordings (not specified)	Lightweight LSTM-based multimodal fusion	83.86% accuracy; model size 0.52 MB	Limited dataset; test on larger, diverse populations
[20]	DAIC-WOZ (Text, Audio, Visual)	RLKT-MDD: Representation learning + knowledge transfer	Accuracy =90%–92%; higher F1 than baseline multimodal models	High complexity; single dataset; extend to larger datasets and real-time clinical use
[10]	**D-Vlog**(961 YouTube vlogs, Audio + Visual)	Multimodal DL with **Cross-Attention**	Accuracy =86%–89% AUC > 0.88, outperforming baselines	No text modality; scalability issues
[21]	MODMA, DAIC-WOZ	EEG-interview multimodal transformer	4.7% accuracy, 10% precision over existing models	Larger datasets and EEG reduction are needed
[22]	Two Chinese multimodal depression datasets (audio, text, video).	IIFDD with intra-modal + inter-modal fusion using attention.	State-of-the-art accuracy =91%–94% on both datasets	Language limitation: IoMT deployment
[23]	1911 audio video text samples 240 balanced subjects used.	AVTF-TBN model using video, audio, text branches, and MMF fusion.	Best F1 = 0.78 using all modalities	Data imbalance; transcription errors

[24]	Multimodal clinical **EEG time-series** and **MRI neuroimaging** data collected from patients	**(HADST)** for EEG–MRI fusion **(CRRCT)**	Accuracy =88%–91% in depression classification	High computational complexity and reliance
[1]	Clinical interview audio recordings (depressed patients)	Vocal prosody analysis (pauses, pitch) and modeling	60% variance predicted, 69% accuracy in severity levels	Test on larger samples, add more vocal features
[25]	Interview-based depression datasets and one synthetic interview dataset.	Multimodal Transformer with adversarial learning	Accuracy =85%–90% improved fairness-aware F1 over baselines	Evaluated only on interview-based datasets and real clinical deployment.

The LLM-based speech content analysis for multilingual depression detection in clinical and general populations. Speech transcripts of Chinese clinical, Italian clinical, and French general population were analyzed. An LLM-based model with few-shot prompting was compared with audio and text embedding for detecting depression and other symptoms like anxiety, insomnia, and fatigue. Result shows that LLM achieved excellent depression detection with an F1-score of 0.96 Chinese, o.85 Italian and 0.40 French outperforming audio and text embedding[11]. Findings show that LLM-based speech analysis provides multilingual capabilities for depression detection without requiring language-specific training data. LLM used to screen depression and anxiety through AI-Generated clinical interviews. Emoscan is used for the early detection of depression and anxiety. The dataset used for this purpose is called psyinterview, with 1157 psychiatrist client dialogues. Emoscan LLM model trained on the psyinterview dataset to classify emotional disorders. Model tested on the dataset and compared with models like GPT4, LIMA3, and Mistral. Emoscan achieved a higher F1-Score of 0.7467, which is better than other LLMS[26]. The state-of-the-art models used in the experiments to achieve optimal outcomes, leveraging multiple modalities. The 275 semi-clinical interviews, including audio, video, and transcripts, were taken from E-DAIC from the AVEC2019 challenge. Model LSTM used for audio and video features, ROBERTa and LLMS (GPT 3.5/4, Lama-3) for text, with Whisper for transcript and audio embeddings. The results showed that LLM (GPT4) achieved RMSE=3.975, MAE=3.16, and CCC=0.781, and classification accuracy of GPT4 and Lama-3 was 71.43 and 73.2%. Further audiovisual multimodal network predicts PHQ-8 scores with RMSE of 6.51, so

text embedding gives the strongest results, especially with LLMS[12]. An automated approach to predicting depression severity using the E-DAIC dataset, in which an LLM model is used for text, and a Bi-LSTM is used for facial features. Three approaches were evaluated: text-based features, facial features, and a combination of both. Methods employed transcript transformation using GPT3.5 Turbo-0125 and fine-tuned DepR oBERTa model combined with question-based features extraction. The findings showed that text data enhancement with speech quality assessment with MAE of 2.85 and an RMSE of 4.02. Further showed that text only with speech quality enhancement performed best[27]. A pipeline for large language models to improve the performance of depression prediction models through the generation of synthetic data. Synthetic data generated from the DAIC-WOZ dataset, which consists of transcripts of interviews between participants and visual interviews. Model used Llama3.2-3B, an open source LLM model from metaAI and BERT trained on original synthetic and combined data. In the first step, the model generated the synopsis and sentiment analysis based on the original transcript, and in the second step, the model generated a synthetic synopsis based on the generated summaries. BERT model performed best with the lowest RMSE 4.64 and MAE 3.66[13].

The effectiveness of LLM (Large Language Models for cost-effective, multilingual depression detection and severity assessment using clinical interview data. Firstly, the performance of four LLMs (GPT4o, GPT4o-mini, Deepseek-v3, Deepseek-R1) was evaluated, and then the best-performing model was selected and tested in a severity and knowledge-enhanced scenario. Four models in zero and few-shot settings were tested on the dataset Doc-Woz (112 interviews), CMDC (78 interviews), and DDSA (51074 statements). The result shows that Deepseek-v3is most reliable in both zero and few-shot scenarios, but zero shot most efficient choice[28].

Retrieval augmented generation (RED) framework used for explainable depression detection. Dataset used DAIC-WOZ with clinical interview transcripts, where labels were assigned using PHQ-8.RED extracts interview information and then retrieves knowledge base evidence. RED achieved the best performance rather than the neural network and LLM baseline on the DAIC-WOZ dataset with an F1 score of 90%[29].

To generate safe CBT (Cognitive behavioral Therapy) based responses using knowledge-infused LLMS. The dataset used the Extended D-Vlog Dataset, which consisted of 1261 real-world YouTube vlogs with multimodal signals. Model Textless language Transformer (TVLT) for multimodal depression detection assessment, depression classification, and CBT response

generation. The results showed that TVLT achieved 67.8% F1-score for depression detection, Mistral achieved 70.1% distortion assessment, 30.9% depression classification, and 88.7% BERT Score for therapeutic responses[30]. Three different NLP models, including BERT, GPT3.5, and GPT-4, on three different datasets for depression detection. Dataset used: DAIC-WOZ, Extended-DAIC, simulated dataset, and KID dataset. BERT was fine-tuned on the DAIC-WOZ dataset, and GPT3.5 and GPT4 were not fine-tuned; they were used in zero-shot or few-shot. GPT4 achieved the highest accuracy, precision, and recall, outperforming both BERT and GPT3.5, even without fine-tuning [31]. AI-powered tools like ChatGPT have the potential to provide valuable mental health support to patients. The semi-structured interviews were conducted in a hospital. Twenty-four patients participated in the interviews and used ChatGPT for mental health support. The results showed that eight positive factors and four negative factors were associated with the use of ChatGPT. Nearly 80% observed accuracy and reliability issues with ChatGPT in mental health [32].

References	**Dataset**	**Method**	**Results**	**Limitation/Future work**
[11]	Chinese (52), Italian (116), and French (1,347) speech data	LLM-based speech content analysis with few-shot prompting	F1-scores: 0.96 (Chinese), 0.85 (Italian), 0.40 (French), high sensitivity in clinical data	Performance varies by language; improve generalization
[26]	PsyInterview 1,157 synthesized clinical dialogues	EmoScan LLM for depression and anxiety screening	F1 = 0.7467 BERTScore = 0.940 Robust external generalization F1 = 0.67	Synthetic dataset; future work: test on larger real-world clinical data
[12]	E-DAIC (275 interviews)	Bi-LSTM (audio/video) + Whisper + RoBERTa + GPT-4/LLaMA-3	GPT-4 best (RMSE 3.98, 71% accuracy)	Small data no raw video need larger clinical datasets & multimodal LLMs)
[27]	E-DAIC (transcripts + videos)	LLM text modeling + Bi-LSTM facial fusion	Text-only accuracy=70%–72%; multimodal fusion improves to =74%–76% accuracy	A few severe cases need more balanced multimodal datasets.
[13]	DAIC-WOZ + synthetic transcripts	BERT (trained on original, synthetic, and combined data)	Best RMSE 4.64; MAE 3.66; improved privacy	Validate on diverse and multimodal data

[28]	51,074 statements from 6 mental disorders	LLM DeepSeek-V3	AUC > 0.90 for most disorders estimated accuracy =85%–90% in zero-shot settings	Poor mild severity detection improves severity & reduces bias.
.[29]	Clinical interview transcripts	RED: Retrieval-Augmented Generation + LLM	Accuracy =82%–87%; higher F1 than NN and standalone LLM baselines improved interpretability	Depends on retrieval quality; scalability
[30]	Extended D-Vlog (1,261 multimodal vlogs)	TVLT + Mistral LLM for detection and CBT responses	F1 = 67.8% (TVLT); 88.7% BERTScore for therapy	Risk of harmful or inaccurate LLM responses. expand real-world datasets
[31]	DAIC-WOZ, Extended-DAIC, KID dataset	BERT (fine-tuned), GPT-3.5, GPT-4	GPT-4 best: accuracy =80%–85%, F1 =0.82; BERT =72% accuracy; GPT-3.5 =75%	Small dataset, add multimodal + larger real-world data
[32]	20 participants from a public hospital using ChatGPT for mental-health support	2-week ChatGPT use + interviews	Qualitative outcomes: >80% users reported that emotional support and CBT guidance were helpful	Ethical risks; improve safety and guidelines

Early detection of depression by using black box machine learning models. Social media user-generated text related to depression was used as a dataset. Convert the social media text into features by using methods such as TF-IDF, Bag of words, N-gram, LDA, and Glove embeddings. Black box models, including SVM, Random Forest, XGBoost, and ANN, were tested on the dataset for depression classification. LIME was used to interpret model predictions. The result shows that SVM achieved the highest accuracy in detecting depression from social media text[33]. Different machine learning classifiers are used to detect whether a person is depressed or not. The mental health assessment dataset used contains 604 participants who were collected through a questionnaire. Extracting features from the dataset, feature selection methods such as SelectkBest, mRMR, and the Boruta feature selection algorithm were used. Classifiers KNN, AdaBoost, GB, XGBoost, Bagging, and Weighted voting were applied. The result showed that the AdaBoost classifier with the SelectkBest feature selection technique is the best model with an accuracy of 92.56%[34]. To identify depression by using higher-order spectral features and fusion features, traditional machine learning and deep learning algorithms were used on speech-related features. The dataset used for the experiment was DAIC-WOZ of the audio/visual emotion challenge AVEC 2017. Methodology included the speech-related feature fusion method

based on HOSA, which was proposed, and secondly, optimized GMMSVM, KNN, and NN models were used to perform classification experiments on speech-related features of the dataset. The result showed that the CNN with fused achieved the best performance with 85% accuracy, CNN outperformed the traditional model, showing that the higher order spectral fusion with deep learning is effective for speech-based depression detection[2]. EVAdaBoost is a new machine learning algorithm that automatically detects sadness from voice recordings by combining signal-processing techniques. It focuses on several speech patterns, while each of these feature sets will train its own AdaBoost model, noise and redundant data are removed, selecting only the most useful features by means of a quantum-inspired evolutionary algorithm. To increase performance without sacrificing accuracy, unnecessary AdaBoost models are eliminated by evolutionary pruning. The results showed that EVAdaBoost outperforms state-of-the-art depression detection algorithms in terms of accuracy, sensitivity, and precision. It has been shown that speech-based mental health screening can be significantly improved by combining evolutionary algorithms[35].

A multimodal feature extraction and decision-level fusion approach is used for the detection of depression with the use of a machine learning algorithm. The database used DAIC-WOZ, so features were extracted from this dataset. The model used SVM and a neural network as a classifier on visual, audio, and text features. GMM (Gaussian mixture model) clustering and Fisher vector were calculated on the relative distance of the facial region. The results showed that this model outperformed the baseline with 17% on audio and 24.5% on visual features[3]. A rich audio dataset DEPAC for mental health, which is labelled with scores on standard scales of two highly prevalent mental disorders. PHQ-9 scores for depression and GAD-7 scores for anxiety assessment. Model used Baseline ML models, including SVM and regression-based classifiers. The results showed that models trained on DEPAC achieved superior depression severity Prediction as compared to other well-known speech corpora[5].

Arab Twitter users express depressive emotions. Researchers collected over 200,000 tweets and carefully examined 1,700 of them after compiling an Arabic dictionary of concepts related to emotions and melancholy. Using tools like Voyant and TAGS, they looked at themes, word frequency, and emotional patterns. They found that people most commonly utilized first-person pronouns, religion (especially the word "Allah"), and dread when talking about sadness. Strong or medium emotional intensity was also evident in several tweets. The findings imply that Arab

users show depression in a different way than English speakers, particularly because of linguistic and religious effects. The findings lay the groundwork for further research on depression identification in Arabic social media[36]. Postpartum depression detection (PPD) using generative AI through free space analysis. The dataset used for detecting PPD was a synthetic dataset (800 users) and a public PPD survey dataset (1491) records. Models include Adaptive random forest, Logistic regression, and Naïve Bayes were tested. Adaptive Random Forest achieved an accuracy of 91% and >90% across all evaluation metrics, so this model outperformed in PPD detection[37]. Google trend search for AI and mental health, AI and depression, AI and anxiety, and focus on public interest and awareness. Web queries were analyzed that are made in search engine with in United States. The Box-Jenkins time series modeling estimation method was used to determine the search volume of AI and mental health by the end of 2024. Results showed that public interest in AI for mental health increased in 2023 and rose by 114% in 2024, revealing a growing awareness and demand for AI-powered mental health solutions [38].

The potential of GenAI tools in transforming mental health care through early detection. A hybrid approach combining a convolutional neural network (CNN), a long short-term memory (LSTM), and transfer-based models to process multimodal data sources. Data used for experiments are social media platforms, wearable sensors, and electronic health records. In testing on 10000 participants, 91% accuracy for depression, 87% in detecting early anxiety symptoms, and 78% of finding the tool useful[39]. Facial action units (AUs) and emotions were biomarkers for depression. Facial expressions were analyzed from video data of participants. Method including feature extraction, mean intensity, comparison of key AUs, and the application of time series classification models. Principal component analysis (PCA) was employed to reduce the dimensionality of AU intensity data. Time series classification applied to facial expression data shows significant differences in the intensity of specific AUs between depressed and healthy ones. The results showed that sadness and happiness were the predominant emotions, highlighting the potential of facial analysis in depression assessment[6]. Depression severity is predicted either as a classification or a regression task, ignoring the ordinality of depression scores. Two synthetic datasets and DAIC-WOZ were used for the experiment. The model used an ordinal regression algorithm for ordinal response data by comparing with multiclass classification and regression using a support vector framework. A further method for

rank boundary estimation is used in RankSVM. The result showed that the method outperforms the existing rank boundary estimation algorithm[40].

References	Dataset	Method	Results	Limitation/Future work
[33]	Social media user-generated text	SVM, RF, XGB, ANN + NLP (TF-IDF, LDA, BoW, GloVe) + XAI (LIME)	Best accuracy =90% with SVM LIME improved interpretability	Limited to social media content, expand datasets, and enhance clinical applicability
[34]	Socio-demographic & psychosocial depression dataset	6 ML classifiers + feature selection SelectKBest, mRMR, Boruta + SMOTE	AdaBoost + SelectKBest achieved the highest accuracy **92.56%** strong sensitivity	Needs clinical validation, expand dataset & include multimodal behavioral features
[2]	AVEC 2017 Depression Sub-Challenge (Speech)	HOS-based Fusion Features + SVM, KNN, CNN	Best accuracy **85% (CNN)** using fused features	Single-modality speech-only limited dataset, extended to multimodal data
[35]	Speech datasets with 9 voice-feature types.	EVAdaBoost evolutionary AdaBoost, BLS learners, CNN time-frequency features.	Accuracy =88%–90%, better than ML baselines	Needs larger datasets; limited to voice future multimodal and real-world testing.
[3]	DAIC-WOZ (AVEC 2017)	SVM, Neural Networks (with decision-level fusion)	Improved over baseline: +17% on audio, +24.5% on visual features on validation set.	Small dataset, handcrafted features, use deep learning-based multimodal fusion for better performance.
[5]	DEPAC (speech-only dataset with multiple tasks and demographics	Baseline ML models (e.g., SVM, regression-based classifiers)	Models trained on DEPAC achieved superior depression severity prediction accuracy =78%	Speech-only modality, baseline models, incorporate deep learning, multimodal data,
[36]	200k plus Arabic tweets (1,700 labeled)	Arabic lexicon, tweet collection, manual annotation, text analysis	Key themes: fear, religion, strong emotions. Classification accuracy 75%	Expand the dataset to include clinically diagnosed users
[37]	Postpartum free-speech dataset (PPD screening interviews)	NLP + ML (tree-based models) + LLM for explainable predictions	90% detection accuracy; real-time, interpretable screening solution	Needs larger, diverse real-world clinical validation, and expand

				multimodal data for robustness
[38]	Google Trends (US, 2023)	Box–Jenkins time series modeling	AI interest in mental health rising; +114% predicted by 2024	Only public search data, clinical impact not assessed
[39]	Social media posts, wearable device data, and electronic health records (US)	Generative AI framework using NLP, sentiment analysis, and behavioral data analytics	Early signs of depression, with an accuracy =83%, and anxiety can be detected from digital footprints for timely intervention	Requires real-world testing and validation for accuracy
[6]	Video data of participants	AU extraction, PCA, time series classification	Sadness & happiness are key to depression detection Accuracy =84% using facial cues	Limited to video; needs multimodal validation
[40]	Synthetic + DAIC-WOZ	Ordinal regression, RankSVM	Outperformed existing rank boundary =10% improvement over baselines =78% accuracy	Focused on ordinal data, extend to multimodal or real-world data

3. Dataset

The following datasets provide a diverse resource base for depression studies and mental health analysis using computational approaches. They include a variety of languages (English, German, Chinese, and multilingual), a variety of modalities (text, audio, visual, EEG, and time series), and established clinical labels and scales such as PHQ-8, PHQ-9, BDI-II, SDS, and GAD-7. (English, German, Chinese, and multilingual), a variety of modalities (text, audio, visual, EEG, and time series), and established clinical labels and scales such as PHQ-8, PHQ-9, BDI-II, SDS, and GAD-7. Benchmark datasets such as DAIC-WOZ, E-DAIC, and the AVEC series provide clinically informed interview data, while text- and social media-driven datasets such as CLPsych 2015, Bell Let's Talk Twitter, DDSA, and Doc-WOZ provide real-world linguistic expressions of depressive symptoms. text- and social media datasets such as CLPsych 2015, Bell Let’s Talk Twitter, DDSA, and Doc-WOZ provide real-world linguistic expressions of depressive symptoms. Video resources such as D-Vlog and Extended D-Vlog facilitate the analysis of verbal and non-verbal behavioral cues, whereas datasets such as MODMA and Google Trends

incorporate physiological signals and population-level behavioral patterns. Video resources such as D-Vlog and Extended D-Vlog facilitate the analysis of verbal and non-verbal behavioral cues, whereas datasets such as MODMA and Google Trends incorporate physiological signals and population-level behavioral patterns. In summary, these datasets complement each other and facilitate the development and evaluation of generative AI-based multilingual and multimodal frameworks for depression detection, while also emphasizing practical considerations with respect to data accessibility, ethics, and privacy.

Dataset Name	Language	Modalities	Subjects No.	Patients No.	Label / Scale	Links
DAIC-WOZ	English	Audio, Visual, Text	189	56	PHQ-8	https://dcapswoz.ict.usc.edu
E-DAIC (AVEC 2019)	English	Audio, Visual, Text	275	–	PHQ-8	https://dcapswoz.ict.usc.edu
AVEC 2013	German	Audio, Visual	292	–	BDI-II	http://avec2013-db.sspnet.eu (access requires registration)
AVEC 2014	German	Audio, Visual	84	34	BDI-II	https://avec2014.dbss.jp/
AVEC 2017	English	Audio, Visual	–	–	PHQ-8	available on request (often bundled with AVEC 2017/E-DAIC)
CLPsych 2015	English	Text	–	–	Binary Depression Label	https://www.cs.jhu.edu/~mdredze/clpsych-2015-shared-task-evaluation/
Bell Let's Talk Twitter	English	Text	–	–	Depression / Control	https://developer.twitter.com/
DEPAC	English	Audio	–	–	PHQ-9, GAD-7	available on request (often research-restricted)
D-Vlog	English	Audio, Visual	816	–	Depression Labels	https://sites.google.com/view/jeewoo-yoon/the dataset
Extended D-Vlog	English	Audio, Visual, Text	1,261	–	Depression Severity	available on request (may require contacting dataset authors)
MODMA	Chinese	Audio, EEG	78	26	PHQ-9, GAD-7, PSQI	https://arxiv.org/abs/2002.09283
EATD-Corpus	Chinese	Audio, Text	162	30	SDS	available on request (as dataset is referenced in

						research)
Doc-WOZ	English	Text	112	–	Depression Severity	https://dcapswoz.ict.usc.edu/
DDSA	Multilingual	Text	51,074	–	Mental Disorder Labels	available on request (text corpus datasets often require an application)
Google Trends (Mental Health)	English	Time-Series	–	–	Search Volume Index	available on request (Google Trends interface)

4. Discussion

4.1 Principal Findings

This paper clearly indicates a transition from machine learning techniques to deep learning approaches for detecting depression using multimodal or generative frameworks. Various pioneering papers on speech features, facial expressions, or text have already validated the usability of the individual modality for detecting depressive symptoms like slowed speech rate, frequent pause durations, depressed linguistic style, and restricted facial expressions. However, the lack of heterogeneity and complexity in unimodal models has often led to the adoption of multimodal models. The fusion techniques based on transformers and cross-attention by more recent approaches have been shown to outperform unimodal and bimodal models. These results immediately support the first objective of this review, to analyze unimodal and multimodal approaches and techniques, and to investigate the role that fusion techniques have on improving the results.

4.2 Contribution of Generative AI and Multilingual Capabilities:

Among the important changes that have been observed in this literature review is the incorporation of large language models (LLMs) in the depression identification system. In this case, several papers have shown the capability of the use of LLMs to detect depression from speech transcripts and interactions for both multilingual and few-shot learning datasets for the Chinese and Italian languages through few-shot prompting. This is even more true when it comes to low-resource languages. There is a lack of labeled data available. Moreover, data generation methods using the LLM have shown success in making the model more robust with the help of generated data that is synthetic yet relevant and can tackle concerns related to data imbalance and privacy issues. Explainable platforms, such as retrieval augmentative generation capabilities,

further promote clinical trust because they can trace their predictions back to pertinent evidence found in the interviews. Such developments are very much on point with the review's focus on multilingual, explainable, and few-shot depression recognition. This is because generative AI technology can be seen as a depression diagnostic booster and data enrichment tool.

4.3 Research Gaps, Clinical Challenges, and Review Contributions:

Although the performance has been promising, various issues are currently impeding practical applications. The majority of the current data sets are small and interview-driven and are demographically constrained; hence, the risk of cultural bias and applicability to platforms such as social media and smartphones currently exists. In particular, the many tasks of predicting the severity of depression can be oversimplified as classification or regression tasks without considering the ordinal characteristics of composite scales like PHQ-8 or BDI-II. In addition, the ethical risks of hallucinated hypotheses, incongruent treatment outcomes, and the lack of regulatory approval must also be considered for the application of generative AI systems in mental health contexts. These obstacles directly led to the key focus of this review, namely, harmonizing data sources, modalities, fusion approaches, and LLM-based approaches with the points of attenuated attention to data paucity, fairness, interpretability, and practicability. In combination with advances into generative AI systems, this review points out the imperative for novel architectures that must move towards multilingual applicability, regulatory approval, interpretability, and alignment with ethical principles focused on effective depression detection and tracking.

5. Conclusion

The present review clear that there was a huge transition from traditional machine learning and unimodal deep learning to multimodal and generative methods based on AI for the identification and detection of depression. Although certain information, like speech and facial expressions related to depression, and sometimes even texts, can be considered through a single modality, these modalities do not usually carry the complexity of the depressive disorder. Fusion of different modalities and, more particularly, the use and application of transformer and cross-attention mechanism-based architectures have demonstrated consistent improvement.

These findings are a contribution to the objective of the review that demonstrates that effective fusion techniques are significant in order to promote robustness and diagnostic accuracy. Big language models have also introduced opportunities to the multilingual, few-shot, and explainer-

based depression identification techniques. Transcript analysis by employing LLMs and artificial data generation is effective in solving the data scarcity and privacy challenges, while explainer-based models improve the interpretability of the model. Some challenges are yet to be completely resolved, such as the lack of diversity in the language data sets used, cultural bias, potential risks to ethics, and the lack of an ordinal scale to assess the severity of the depression. The researcher must focus on the multi-modal, large-scale data sets that are culturally diverse and are aware of the ordinals to create a certain element of the clinical practice credible so that it reflects real-world scenarios.

References

[1]. Y. Yang, C. Fairbairn, J. F. Cohn, and A. Member, “from Vocal Prosody,” pp. 1–9.

[2]. X. Miao, Y. Li, M. Wen, Y. Liu, I. N. Julian, and H. Guo, “Fusing features of speech for depression classification based on higher-order spectral analysis,” vol. 143, no. July, pp. 46–56, 2022.

[3]. S. Dham, A. Sharma, and A. Dhall, “Depression Scale Recognition from Audio, Visual and Text Analysis,” 2016.

[4]. A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, “Deep Learning for Depression Detection of Twitter Users,” pp. 88–97, 2018.

[5]. M. Tasnim, M. Ehghaghi, B. Diep, and J. Novikova, “DEPAC : a Corpus for Depression and Anxiety Detection from Speech,” vol. 2013, pp. 1–16, 2022.

[6]. A. Parikh and M. Sadeghi, “Exploring Facial Biomarkers for Depression through Temporal Analysis of Action Units,” vol. 1, 2018.

[7]. L. He, “DepNet : An Automated Intelligent System using Deep Learning for Video-based Depression Analysis,” pp. 1–24.

[8]. N. Marriwala and D. Chaudhary, “Measurement : Sensors,” vol. 25, no. September 2022, pp. 0–9, 2023.

[9]. C. Xu *et al.*, “Deep learning-based detection of depression by fusing auditory , visual and textual clues,” vol. 391, no. July, 2025.

[10]. J. Yoon, C. Kang, S. Kim, and J. Han, “D-vlog : Multimodal Vlog Dataset for Depression Detection”.

[11]. R. Riad, A. Ducorroy, and A. Lesage, “Automated speech content analysis to detect depression with large language models : towards multilingual and few-shot capabilities,” pp. 1–23, 2025.

[12]. C. Tank, S. Pol, V. Katoch, S. Mehta, A. Anand, and R. R. Shah, “Depression Detection and Analysis using Large Language Models on Textual and Audio-Visual,” pp. 1–12.

[13]. A. Kang, J. Y. Chen, Z. Lee-youngzie, and S. Fu, “Synthetic Data Generation with LLM for Improved Depression Prediction,” 2022.

[14]. K. Daly and O. Olukoya, “Biomedical Signal Processing and Control Depression detection in read and spontaneous speech : A Multimodal approach for lesser-resourced languages,” vol. 108, no. May 2024, 2025.

[15]. A. Y. Kim, E. H. Jang, S. Lee, and K. Choi, "Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals : Deep Convolutional Neural Network Approach Corresponding Author :," vol. 25, pp. 1–17, 2023, doi: 10.2196/34474.
[16]. N. Seneviratne and C. Espy-wilson, "Speech based Depression Severity Level Classification Using a Multi-Stage Dilated CNN-LSTM Model," pp. 0–4.
[17]. E. Lim, M. Jhon, J. Kim, S. Kim, S. Kim, and H. Yang, "A lightweight approach based on cross-modality for depression detection," vol. 186, no. January, 2025.
[18]. M. Rojc, "An End-to-End framework for extracting observable cues of depression from diary recordings," vol. 257, no. December 2023, 2024.
[19]. L. Zhang and S. Zhang, "A Multimodal Artificial Intelligence Model for Depression Severity Detection Based on Audio and Video Signals," 2025.
[20]. S. Yang, L. Cui, L. Wang, T. Wang, and J. You, "Heliyon Enhancing multimodal depression diagnosis through representation learning and knowledge transfer," vol. 10, no. January, 2024.
[21]. N. Esmi, "Biomedical Signal Processing and Control Multimodal transformer for depression detection based on EEG and interview data," vol. 113, no. November 2025, 2026.
[22]. J. Chen, Y. Hu, Q. Lai, W. Wang, J. Chen, and H. Liu, "IIFDD : Intra and inter-modal fusion for depression detection with multi-modal information from Internet of Medical Things," vol. 102, no. July 2023, 2024.
[23]. G. Huang, Z. Liang, L. Zhang, L. Li, H. Ding, and Z. Zhang, "Multimodal Sensing for Depression Risk Detection : Integrating Audio , Video , and Text Data," 2024.
[24]. J. Li, Y. Chen, and J. Wang, "Intelligent assessment method for mental disordersby integrating EEG and MRI multimodalhigh-dimensional data Intelligent assessment method for mental disorders by integrating EEG and MRI multimodal high-dimensional data," pp. 0–21, 2025.
[25]. E. Zhang and C. Poellabauer, "Mitigating Interviewer Bias in Multimodal Depression Detection : An Approach with Adversarial Learning and Contextual Positional Encoding," pp. 12169–12188, 2025.
[26]. J. M. Liu and M. Gao, "Enhanced large language models for effective screening of depression and anxiety," pp. 1–11, 2025.
[27]. M. Sadeghi *et al.*, "Harnessing multimodal approaches for depression detection using large language models and facial expressions," 2024.
[28]. L. Xian, J. Ni, and M. Wang, "Leveraging Large Language Models for Cost-Effective , Multilingual Depression Detection and Severity Assessment".
[29]. R. Generation, "Explainable Depression Detection in Clinical Interviews with Personalized Retrieval-Augmented Generation," 2022.
[30]. P. Moon and P. Bhattacharyya, "We Care: Multimodal Depression Detection and Knowledge Infused Mental Health Therapeutic Response Generation," 2018.
[31]. B. Hadzic *et al.*, "Enhancing early depression detection with AI : a comparative use of NLP models Enhancing early depression detection with AI : a comparative use of NLP," vol. 4889, 2024, doi: 10.1080/18824889.2024.2342624.
[32]. "Mental health support.pdf."
[33]. S. Hameed, M. Nauman, N. Akhtar, M. A. B. Fayyaz, and R. Nawaz, "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," no. September, pp. 1–19, 2025, doi: 10.3389/frai.2025.1627078.
[34]. "ML to predict depression.pdf."

[35]. R. Sayeri, B. Barzegar, Y. Bozorgi, and N. Mikaeilvand, “Machine Learning with Applications Evolutionary AdaBoost ensemble : A machine learning framework for depression detection,” vol. 22, no. October, 2025.

[36]. A. Mohamed and W. Zaghouani, “ScienceDirect Expression of Depression Among Arab Twitter Users Using Arabic Corpus Analysis,” vol. 244, pp. 76–85, 2024.

[37]. S. García-méndez, F. De Arriba-pérez, S. García-méndez, and F. De Arriba-pérez, “Detecting and Explaining Postpartum Depression in Real-Time with Generative Artificial Intelligence Detecting and Explaining Postpartum Depression in Real-Time with Generative Artificial Intelligence,” *Appl. Artif. Intell.*, vol. 39, no. 1, 2025, doi: 10.1080/08839514.2025.2515063.

[38]. S. Banerjee, P. Dunn, S. Conard, and A. Ali, “Mental Health Applications of Generative AI and Large Language Modeling in the United States,” 2024.

[39]. S. Kumar, “Early detection of depression and anxiety in the USA using generative AI,” vol. 7, no. 1, pp. 1–7, 2025.

[40]. S. Jayawardena, “Support Vector Ordinal Regression for Depression Severity Prediction,” no. 2, pp. 3–8.