Frontiers in Psychiatry

TYPE Review
PUBLISHED 13 November 2024
DOI 10.3389/fpsyt.2024.1417253

An historical overview of artificial intelligence for diagnosis of major depressive disorder

Hao Liu[1,2†], Hairong Wu[1†], Zhongli Yang[1], Zhiyong Ren[3], Yijuan Dong[2,4], Guanghua Zhang[5*] and Ming D. Li[1*]

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2]Shanxi Tongchuang Technology Inc., Taiyuan, China, [3]Shanxi Province Mental Health Center, Taiyuan Psychiatric Hospital, Taiyuan, China, [4]Shanxi Yingkang Healthcare General Hospital, Yuncheng, Shanxi, China, [5]School of Big Data Intelligent Diagnosis and Treatment Industry, Taiyuan University, Taiyuan, China

OPEN ACCESS

EDITED BY
Saïd Boujraf,
University of Fez, Morocco

REVIEWED BY
Zakariae Alami Merrouni,
Sidi Mohamed Ben Abdellah University,
Morocco
Mohamed Merzouki,
Université Sultan Moulay Slimane, Morocco

*CORRESPONDENCE
Ming D. Li
ml2km@zju.edu.cn
Guanghua Zhang
zhangguanghua@tyu.edu.cn

†These authors have contributed equally to the work

RECEIVED 15 April 2024
ACCEPTED 10 October 2024
PUBLISHED 13 November 2024

CITATION
Liu H, Wu H, Yang Z, Ren Z, Dong Y, Zhang G and Li MD (2024) An historical overview of artificial intelligence for diagnosis of major depressive disorder. *Front. Psychiatry* 15:1417253. doi: 10.3389/fpsyt.2024.1417253

The Artificial Intelligence (AI) technology holds immense potential in the realm of automated diagnosis for Major Depressive Disorder (MDD), yet it is not without potential shortcomings. This paper systematically reviews the research progresses of integrating AI technology with depression diagnosis and provides a comprehensive analysis of existing research findings. In this context, we observe that the knowledge-driven first-generation of depression diagnosis methods could only address deterministic issues in structured information, with the selection of depression-related features directly influencing identification outcomes. The data-driven second-generation of depression diagnosis methods achieved automatic learning of features but required substantial high-quality clinical data, and the results were often obtained solely from the black-box models which lack sufficient explainability. In an effort to overcome the limitations of the preceding approaches, the third-generation of depression diagnosis methods combined the strengths of knowledge-driven and data-driven approaches. Through the fusion of information, the diagnostic accuracy is greatly enhanced, but the interpretability remains relatively weak. In order to enhance interpretability and introduce diagnostic criteria, this paper offers a new approach using Large Language Models (LLMs) as AI agents for assisting the depression diagnosis. Finally, we also discuss the potential advantages and challenges associated with this approach. This newly proposed innovative approach has the potential to offer new perspectives and solutions in the diagnosis of depression.

KEYWORDS

major depressive disorder, MDD, artificial intelligence, large language model, AI agent, multimodal diagnosis

1 Introduction

Depression is a prevalent and chronic mental illness that significantly impacts an individual's cognition, emotions, and behaviors (1, 2). With the rapid development of modern society and the increasing pressures of work and life, the global prevalence of depression has reached 4.4%, making it one of the most common and severe mental disorders worldwide (3). MDD affects as many as 12% of adults globally, with the highest prevalence in the United States observed among the youth, women, and elderly (4). In China, the lifetime prevalence of severe depression is 3.4%, with a 12-month prevalence of 2.1% (5). In the context of the escalating incidence of depression, clinical diagnosis primarily relies on scale screening and psychiatric interviews. However, the diagnostic outcomes of this method depend on the clinician's experience and the patient's cooperation, leading to a high rate of misdiagnosis. The Value of Treatment study conducted by the European Brain Council suggested that approximately 52% of MDD episodes went undiagnosed, and about 38% of diagnosed patients did not receive treatment (6). Consequently, there is an urgent need for an objective, efficient, accurate, and convenient diagnostic method for depression.

With the rapid advancement of AI technology, a substantial body of researches has explored the application of AI in assisting the diagnosis and prediction of depression. The first generation of knowledge-driven approaches for depression diagnosis primarily relied on classical Machine Learning (ML) techniques for prediction. These approaches involved the extraction of features from various sources, such as speech signals, facial expressions, and Electroencephalogram (EEG) data, by using feature extraction algorithms. ML models were then constructed based on these extracted features for depression prediction and diagnosis (7, 8). However, the feature extraction process often requires supposed prior knowledge and the loss of information can be inevitable. The research trend has since shifted from traditional handcrafted feature design towards more advanced methods that can automatically learn features. With the increase of clinical data related to depression and the development of Deep Learning (DL) techniques, more and more approaches based on DL models have been applied to depression diagnosis, marking the birth of data-driven second-generation depression diagnostic approaches. DL models, designed end-to-end, have the capability to automatically learn features during model training. They could encode and classify data without relying on feature extractors which designed in advance, thereby automating the process of depression diagnosis (9, 10). However, DL networks remain black-box models, often lacking interpretability. For reasons of safety in the medical industry, relying solely on the diagnostic results generated by DL networks can be challenging for clinical application. The core idea behind the third generation of depression diagnostic methods is to combine the strengths of the previous two generations to enhance diagnostic accuracy. By integrating manually extracted features with those represented by DL models and employing various multimodal feature fusion strategies, a more comprehensive and accurate diagnosis of depression can be achieved. The combination of manually extracted features with complex models can better uncover depressive feature information, while information fusion, by combining heterogeneous information from different sources, ensures high-quality information processing for more comprehensive and reliable decision-making (11). Nevertheless, the third-generation diagnostic approaches still lack of interpretability which are critical in clinical practices. Additionally, while they can significantly improve the diagnostic accuracy, they tend to complicate the overall diagnostic process, moving away from the goal of the fully automated depression diagnosis.

With the proliferation of data and computational power available for training AI models, models like GPT-4 (Generative Pre-trained Transformer 4) (12, 13), Llama (14, 15), and GLM (General Language Model) (16, 17), among others, have achieved significant success. These LLM models exhibit strong generalizability and emergent abilities. Researchers have begun exploring their potential applications in the medical field, including medical knowledge-based question answering, medical X-ray interpretation, and medical treatment recommendations. Given the versatility and efficiency of LLMs and their highly interpretable text-based outputs, combining the goal of automating depression diagnosis with LLMs thus represents a new direction in the realm of depression diagnostic assistance.

In this review, we conducted a systematic literature search across PubMed, IEEE Xplore, and Google Scholar, covering the period from January 2010 to April 2024. The search used combinations of keywords such as "Artificial Intelligence", "Major Depressive Disorder", "MDD", "depression", "diagnosis", "machine learning" and "deep learning". We only included peer-reviewed studies published in English that focused on the application of AI to the diagnosis of MDD and provided detailed methodologies and quantitative results. We excluded those non-peer-reviewed works, studies lacking methodological clarity or quantitative outcomes, or studies focused solely on treatment. Articles were initially screened by title and abstract, followed by a full-text review. A total of 114 studies were selected based on relevance, methodological rigor, and their contribution to the field, ensuring that key advancements and gaps in AI-based depression diagnosis were thoroughly addressed.

2 Scale-based evaluation methods

The assessment of depression using scale-based approaches primarily includes self-assessment of scales and physician assessment according to clinical diagnostic criteria (Table 1). Commonly recognized diagnostic criteria include International Classification of Diseases, 11th (ICD-11) (18), Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-V) (19, 20), and Mini-International Neuropsychiatric Interview (MINI) (21, 22). Other commonly used clinical scales still include Patient Health Questionaire-9 Items (PHQ-9) (23, 24), the Self-Rating Anxiety Scale (SAS) (25, 26), and the Self-Rating Depression Scale (SDS) (27, 28).

The PHQ-9 is currently the most used self-rating scale, where users evaluate their behavior and total scores over the past two weeks based on scoring criteria. The Hamilton Depression Scale-17 Items (HAMD-17) (29) scale is one of the most commonly used ones, which requires a combination of two professional physicians to evaluate patients and independently score them. Generally, a combination of conversation and observation is used. Among them, the self-rating

TABLE 1 Clinical diagnostic criteria and related scales for depression.

Function	Scale	Pattern	Scale Description
Clinical diagnostic criteria	DSM-V	Others-rating	It's a comprehensive guide for diagnosing mental disorders, featuring standardized criteria essential for clinical practice, research, and education in mental health.
	ICD-11	Others-rating	ICD-11 is global classification system by WHO for diseases and health conditions, aiding diagnosis and documentation worldwide.
	MINI	Others-rating	The MINI is a brief psychiatric diagnostic interview used by mental health professionals, covering various disorders efficiently and thoroughly.
Depression scales	PHQ-9	Self-rating	PHQ-9 is a self-administered questionnaire used as a screening tool for depression. It consists of nine questions based on the Diagnostic and Statistical DSM-V criteria for major depressive disorder. PHQ-9 scores help assess the severity of depression symptoms.
	HAMD-17	Others-rating	HAMD-17 is a clinician-administered questionnaire used to assess the severity of depression symptoms in individuals. It comprises 17 items covering various aspects of depression, such as mood, guilt, suicidal ideation, and sleep disturbances.
	SDS	Self-rating	SDS is a self-report depression screening tool developed. It consists of 20 items designed to measure the level of depression in individuals.
	MARDS	Others-rating	MADRS is a clinician-administered tool comprising 10 items to assess depression severity. It evaluates symptoms like sadness, tension, sleep problems, and concentration difficulties, aiding clinicians in monitoring depression and treatment response.
	BDI	Self-rating	BDI is a self-report questionnaire assessing depression severity in adults and adolescents. With 21 items covering mood, guilt, and physical symptoms, it yields scores indicating the severity of depressive symptoms.

DSM-V, Diagnostic and Statistical Manual of Mental Disorders; ICD-11, International Classification of Diseases; MINI, Mini-International Neuropsychiatric Interview; PHQ-9, Patient Health Questionaire-9 Items; HAMD-17, Hamilton Depression Scale; SDS, Self-rating Depression Scale; MARDS, Montgomery-Asberg Depression Rating Scale; BDI, Beck depression rating scale.

scale is relatively simple and convenient, but the answer of users to the scale question solely based on personal self-awareness is more subjective. The other-rating scales used in clinics is relatively time-consuming and laborious, but more professional and accurate. In evaluating depression scales, Wu et al. (30) utilized the Hospital Anxiety and Depression Scale-Depression subscale (HADS-D) to screen for depression. Their results demonstrated that HADS-D exhibited varying sensitivity and specificity depending on the type of diagnostic interview employed. Among 101 eligible studies, which included 25,574 participants with 2,549 diagnosed with MDD, it was found that for semi-structured interviews, fully structured interviews, and the MINI, the optimal cut-off value for HADS-D was 7 or higher, yielding a sensitivity of 0.82 and a specificity of 0.78. As the cut-off value increased, specificity improved, but sensitivity decreased; for instance, at a cut-off value of 11, sensitivity was 0.44, and specificity was 0.95 [Wu et al. (30)]. These findings indicate that HADS-D's performance varies across different diagnostic contexts, and the choice of cut-off value significantly influences the accuracy of screening results. Thus, this suggests that, in clinical practice, different cut-off values should be selected on the basis of specific needs of patient populations in order to balance the risks of false positives and false negatives, thereby more effectively identifying patients with depression.

3 Automated diagnosis approaches combining with AI

Diagnostic results based on traditional scale-based methods rely heavily on the experience of clinicians and patient cooperation, which could lead to a high rate of misdiagnosis. Additionally, limited healthcare resources imply that diagnostic services are probably available to only a small portion of the population. In contrast, AI technology can automate the analysis of respondents' linguistic content and facial expressions, reducing reliance on self-reported data. This approach enhances the objectivity of the assessment, decreases the likelihood of misdiagnosis, and expands the accessibility of diagnostic services.

With the advancement of diagnostic and therapeutic techniques for depression, an increasing volume of research related to depression has generated a substantial amount of clinical diagnostic and treatment data. AI technology, with its inherent strengths in handling large datasets, multidimensional information, and multimodal data holds great promise in this context (31, 32). Therefore, leveraging the automatic learning capabilities of AI to depression diagnosis represents an effective method towards automating the diagnosis of depression (33, 34). In the following sections, we will provide a description of each generation of AI technology in terms of their research and applications on depression.

3.1 The first-generation diagnostic methods for depression

The depression diagnostic model, constructed using traditional ML algorithms, is referred to as the first-generation depression diagnostic method. This method relied on robust prior knowledge and necessitated the use of various feature extraction algorithms to identify depression-related features. Subsequently, ML models with relatively simple structures were built based on these features. This enabled the model to automatically learn from the feature data and optimize its performance, ultimately achieving the goal of assisting in depression prediction.

3.1.1 Text features methods

The linguistic characteristics of depressed patients often manifest as pessimism, low self-esteem, and unclear, repetitive, or incoherent speech content (35). Currently, classification methods on text contents can be broadly categorized into grammar rules combined with statistical approaches and ML methods.

Grammar rules combined with statistical methods typically involve the manual construction of domain-specific sentiment lexicons. During the sentiment recognition process, a comparison of the words in the input text with the domain-specific sentiment lexicon or calculates vector distances is commonly used to identify matching sentiment words and their polarities. In the context of constructing Chinese depression sentiment lexicons, Yin et al. (36) proposed a sentiment word mining method that combines word frequency statistics and sentiment word intensity. Liu et al. (37) expanded upon a basic sentiment lexicon by using the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm to select domain-specific sentiment lexicons and extend them through word similarity calculations. When using sentiment lexicons for text classification, He et al. (38) introduced a PRE-TF-IDF classification algorithm based on weight preprocessing. This method enhances text classification accuracy by adding weight preprocessing and word density weighting to the TF-IDF algorithm.

For traditional ML approaches, the primary method involves manually extracting features from sample datasets and selecting suitable classification models based on these features. The choice of an appropriate classifier based on text features is crucial for text sentiment classification, and commonly used classifiers include Support Vector Machine (SVM), Naive Bayesian (NB), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), among others. Zhang et al. (39) conducted a comparative study on the prediction of depression diagnosis using different ML algorithms based on NHANES data. The results showed that SVM and CatBoost classifiers performed best in the total feature set, while they had the lowest performance in the dietary feature set. From a data application perspective, Diao et al. (40) systematically analyzed several reports on ML-based text classification and concluded that SVM is frequently used in medical datasets due to its ability to effectively overcome challenges such as high-dimensional calculations and overfitting.

3.1.2 Speech features methods

Through clinical observation and statistics, individuals with depression typically show the following characteristics in their speech behavior: slow speech rate, low voice intensity, reduced spontaneity, and unclear articulation, among others (41). These speech behavioral characteristics can be manifested through various speech features, including prosodic features such as pitch, intonation, energy, and rhythm variations, as well as fundamental frequency features resulting from vocal cord vibrations, and speech rate features that express the speed of speech. It is known that individuals with depression usually exhibited a faster decline in speech rate and longer response times (42). There existed a negative correlation between pitch and loudness with the severity of depression (43). Shankayi et al. (44) also found that a combination of features, including prosody, voice spectrum, and glottal features, yields better recognition results than any single feature type. In the realm of audio feature sets, Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (45) and COVAREP are widely recognized feature sets. eGeMAPS is often used as an acoustic feature set in depression detection challenges like the audio/visual emotion challenge and workshop (AVEC) (46).

Sumali et al. (47) collected data from 300 depression patients and used the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm to perform feature selection on the participants' audio data to predict depression using a linear SVM model. They achieved a classification accuracy of up to 93.3% on their test set. Cummins et al. (48) employed Gaussian Mixture Models (GMMs) to model depressive and neutral speech, using Mel Frequency Cepstral Coefficient (MFCC) features from audio as the primary discriminative feature. They tested their approach on a dataset of 23 depression patients and 24 control subjects, achieving a classification accuracy of 82%. Jiang et al. (49) analyzed the classification performance of prosodic, spectral, and glottal speech features in depression recognition and proposed an ensemble logistic regression model, ELRDD-E, for detecting depression in speech. In an automatic depression classification task on a dataset of 170 Chinese native subjects, the accuracy rate for females was 75.00%.

3.1.3 Visual features methods

The mental state of individuals with depression is correlated with changes in their facial expressions. Common features observed in the visual analysis of their faces include a lack of expression, a somber demeanor, a forlorn look, and evasive eye contact (1). For instance, there are statistically significant differences in eye movement patterns between depression patients and healthy individuals (50). Depression is often associated with greater downward gaze angles, shorter average gaze durations, and reduced smile intensity (51). In studies related to psychomotor disturbances in bipolar disorder, depression patients exhibit longer reaction times in gaze tasks (52). In the domain of facial feature extraction tools, the OpenFace package (53) stands as a cutting-edge open-source tool for facial motion analysis. This tool provides functionalities such as facial landmark detection, head pose tracking, eye gaze tracking, and facial action unit recognition.

Wang et al. (54) extracted key facial features from collected facial videos using a human activity appearance model. By using SVM for depression classification based on variations in eye, eyebrow, and mouth movements, the authors achieved an accuracy of 78.85%, a recall rate of 80.77%, and an F1-score of 0.792 in a sample with 26 subjects. Li et al. (55) extracted facial features, including eyes, eyebrows, mouth, and nose, and utilized an average performance of three classification models: SVM, NB, and Random Forest (RF), for predicting depression in both male and female patients, which achieved an average classification accuracy of 71.5% for females and 66.7% for males. Additionally, their results indicated that eyebrows and mouth contribute more to neutral emotional valence than other facial areas. Li and Fan (56) conducted a review of facial feature studies in depression patients from the year of 2016 to 2021 using ML approach. They concluded that SVM, NB, RF are the most commonly used classifiers in facial feature research of depression patients. Further, they pointed out that many studies overlook the temporal aspects of emotional changes in depression patients and suggested that multimodal fusion approach is an important future research direction, particularly in the context of limited clinical sample size and relatively narrow research objectives (57).

3.1.4 Other modal features methods

In addition to the research on behavioral features in depression patients, researchers have also explored various other modalities, including brain imaging, electrophysiological data, and omics data for studying depression. In the first-generation research on brain imaging for depression diagnosis, Bhaumik et al. (58) utilized features from the left posterior cingulate cortex and the right posterior dorsolateral prefrontal cortex. They employed the SVM algorithm to distinguish severe depression patients from healthy controls, achieving an accuracy of 76.1%. Wei et al. (59) extracted resting-state networks and calculated their Hurst index using range-scale analysis. They achieved a recognition rate of 90% for depression patients using a linear kernel SVM. In research on electrophysiological data for assisting depression diagnosis, Akbari et al. (60) proposed a method for detecting depression using brainwave signal reconstruction in phase space and geometric features. The framework, with features selected using particle swarm optimization and a support vector machine classifier, achieved an average classification accuracy of 99.30% and a Matthews Correlation Coefficient (MCC) of 0.98. In the domain of omics data, Dipnall et al. (61) employed ML enhanced regression algorithms and logistic regression to identify three key biomarkers related to depression: red cell distribution width, serum glucose, and total bilirubin.

3.1.5 Multimodal features methods

Differences in feature data for depression patients manifest across multiple modalities, and the information from these modalities complements each other. Therefore, the development of depression diagnosis methods that combine multi-modal data features is an important research direction for achieving automated diagnostic assistance.

Yang et al. (62) combined audio, video, language, and sleep data features and used a decision tree model to distinguish between normal individuals and depression patients. They tested their approach on the Distress Analysis Interview Corpus of human and computer interviews (DAIC) dataset (63), achieving an average F1-score of 0.724. Alghowinem et al. (64) fused audio features, gaze features, and head features of participants and optimized an SVM classifier using Library for Support Vector Machines (LibSVM) (65). Subsequently, in a real clinical validation dataset consisting of 30 patients with severe depression and 30 healthy controls, the accuracy of single-modal classification results was 83% for audio, 73% for gaze, and 63% for head features. The fused multimodal approach achieved an average accuracy of 88%, effectively demonstrating the complementary nature of features across different modalities. Zhao et al. (66) combined speech and facial visual data to propose a multi-modal fusion depression diagnosis algorithm. They introduced spectrum subtraction, orthogonal matching pursuit algorithms, and a cascade based on the ratio of sound and facial emotion. Their approach achieved a diagnostic accuracy of 81.14%, which was 6.76% higher than using only audio as a single modality.

3.1.6 Summary of the first-generation diagnostic methods for depression

The first-generation AI depression diagnosis methods relied on prior knowledge and traditional ML algorithms to identify depression patients. These methods exhibited strong interpretability. In these approaches, the feature selection in ML directly influenced the best performance of the identification results. Therefore, it is essential to extract a sufficient number of depression-related features in order to effectively improve the recognition accuracy. Our current understanding of depression is considered to be limited, which could lead to partial data presentation and, to some extent, restrict the effectiveness of model classification. Consequently, the extraction of a sufficient number of depression-related features formed the foundation for the quality of these methods. Furthermore, the process of manually extracting prior knowledge from raw data was often inefficient. Therefore, the research trend in AI-based depression diagnosis methods has shifted from traditional handcrafted feature design to more advanced methods that involve automatic feature learning. The first-generation framework for depression diagnosis methods is summarized in Figure 1.

3.2 The second-generation diagnostic methods for depression

Inspired by the brain's neural system, specific approaches based on DL models have been successfully applied in the realm of sentiment analysis. One such application is the development of a depression diagnostic model, termed the second-generation depression auxiliary diagnostic method. The primary problem it addressed was how to automate the extraction of various modality feature data from respondents and train a diagnostic model based on a large volume of feature data. In this section, we provide a comprehensive overview of existing second-generation diagnostic methods, categorizing them broadly based on the types of data employed into textual, audio, video, other type, and multimodal data-driven depression auxiliary diagnostic methods.

3.2.1 Text features methods

In the context of depression recognition research that combines social network text data with DL, Xu (67) proposed a depression analysis model based on an attention mechanism. This model takes into account the characteristics of social text data, employs automatic feature selection techniques, and utilizes a multi-

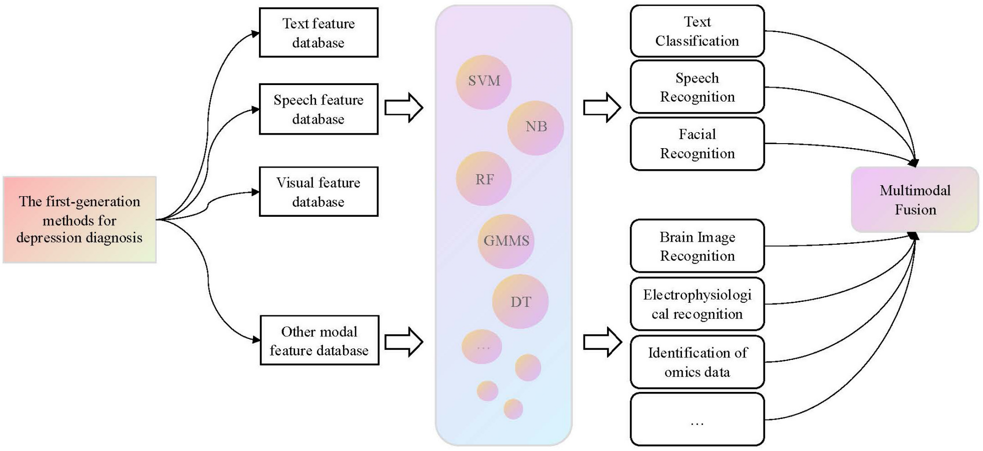


FIGURE 1
The first-generation framework for depression diagnosis methods. For this generation of methods, it is necessary to first extract modal features related to depression from the raw data, and then use classical ML algorithms to classify these features to identify patients with depression.

instance learning approach. Ive et al. (68) introduced an Recurrent Neural Network (RNN) architecture that incorporates an attention mechanism to predict potential depression patients. This model can effectively extract essential textual elements and make accurate predictions, due to its integration of the attention mechanism. Coppersmith et al. (69) conducted research on the presence of suicide attempts within social network text. They used Long Short-Term Memory (LSTM) networks to quantify textual signals related to suicide attempts and captured contextual information between textual contents, thus obtained text content associated with suicide.

3.2.2 Speech features methods

In research focused on diagnostic methods based on acoustic features, Chlasta et al. (70) introduced an ensemble learning approach based on Convolutional Neural Networks (CNN), which exhibited promising performance on the AVEC2016 dataset (71). Lu et al. (72) addressed the issues of complex model structures and low recognition rates associated with Deep Neural Network (DNN) methods and proposed a model that combines residual thinking and attention mechanisms, achieving an accuracy rate of 76%. Li and Fu (73) conducted research from both feature design and network architecture perspectives by introducing a Multi-scale Audio Differential Normalization (MADN) feature extraction algorithm and developed a depression recognition regression model called DR AudioNet based on this algorithm.

3.2.3 Visual features methods

Depression diagnostic methods that combine DL techniques with facial visual features can generally be categorized into two major approaches. One is the "Visual Local" approach, which is based on previous research on facial features associated with depression. In this approach, specific local features related to depression are identified and used as inputs to DL models. This method requires fewer features and tends to converge more easily during model training. The other is the "Visual Global" approach, which aims to prevent information loss during the feature extraction process by treating the entire facial information as input to the model for automated feature extraction. However, this method requires extensive training on a large amount of video data and the construction of complex DL models.

In the "Visual Local" research, feature extraction methods primarily involve using visual extraction tools, integrated systems, or various devices such as eye-tracking devices to directly extract feature information from specific facial regions. These features include changes in facial landmark points, the frequency of muscle unit occurrences, eye gaze angles, and more. For example, Wang et al. (54) achieved a recognition rate of 78% by utilizing features extracted from eye movements, eyebrow movements, and pupil motions obtained during interviews, combined with other modal features. Gao (74) built a depression recognition model based on multi-scale facial features using facial Action Units (AUs) and facial landmark points and demonstrated that multi-scale features outperformed single-scale features.

In "Visual Global" research, the entire video is typically used as input to the model for automatic feature extraction. Xu (75) addressed the varying impact of different facial regions on the recognition of depression and constructed the dual pooling residual attention network (DPRAN), incorporating a residual attention mechanism. This network was used to extract spatiotemporal features from facial images to complete the depression recognition task. De Melo et al. (76) explored global features, local features, and spatiotemporal correlations of facial regions and proposed a depression recognition architecture based on 3D convolutional neural networks.

3.2.4 Other modal features methods

In the research on depression auxiliary diagnosis methods based on EEG, Acharya et al. (77) employed a CNN model for depression screening on the basis of EEG signals. This method allows for automatic adaptive learning from input EEG signals to differentiate between EEG signals of individuals with and without depression. The algorithm achieved recognition accuracies of 93.5% and 96.0% for left and right hemisphere EEG signals, respectively, in a sample size of 30 subjects. In the study of predicting antidepressant response to severe depression using genetic biomarkers, Lin et al. (78) developed a DL predictive method, with the goal of predicting the possible antidepressant treatment outcomes for MDD using 10 Single Nucleotide Polymorphism (SNP) data associated with antidepressant response. The predictions were made using a Multi-layer Feedforward Neural Network (MFNN) model with two hidden layers, achieving the Area Under the ROC Curve (AUC) of 0.8228 and sensitivity of 0.7546.

3.2.5 Multimodal features methods

Previous research has indicated that methods based on single-modal features such as text content, audio signals, and facial visual cues have produced promising results in depression diagnosis. However, the behavioral characteristics of individuals with depression manifest in multiple modalities, and the information across these modalities is complementary. Results from AVEC (46) have consistently demonstrated that, compared to single-modal depression recognition methods, multimodal fusion methods tend to yield a better performance. The framework for second-generation depression diagnostic methods is depicted in Figure 2.

In the realm of research on multimodal fusion methods in second-generation diagnostic approaches, Wang et al. (79) summarized nine statistical features related to user text, social behavior, and images as well. They extracted deep representations of text using the XLNet model and integrated them with other modal data to construct a Multimodal Feature Fusion Network (MFFN). This method achieved an F1-score of 0.9685 on the WU3D test dataset. Hao (80) employed Chinese Bidirectional Encoder Representations from Transformers (ChineseBERT), ResNet101 V2, and Bidirectional Long Short-Term Memory (BiLSTM) models to extract features from text, visual information, and audio data, respectively, and used an attention algorithm to fuse these features and effectively calculated the complementarity and contribution of multimodal information by continuously adjusting the weights of each modality.

3.2.6 Summary of the second-generation diagnostic methods for depression

The second-generation depression diagnostic methods typically utilize different DL models trained on large-scale datasets as tools for modality feature representation. In contrast to first-generation diagnostic methods, they do not require prior knowledge and circumvent the complex process of manual feature extraction. Through an end-to-end design, these methods can automatically learn feature extraction and achieve depression diagnosis (9, 81). However, up to this point, researchers are still unable to explain the inner workings of DL networks. With DL networks, they can only obtain diagnostic results without insight into the basis for the diagnosis. Since DL networks are considered to black-box models, relying solely on the diagnostic results from deep networks may pose challenges in clinical applications, especially concerning medical industry safety-related issues.

3.3 The third-generation diagnostic methods for depression

The third-generation diagnostic method for depression combines the advantages of the first and second-generation methods, aiming to improve the accuracy of depression diagnosis. It utilizes manually extracted depression features as prior knowledge and also employs deep learning models to capture deep representation information from different modalities of features. Finally, it integrates multiple sources of information through various fusion strategies to achieve depression recognition. In terms of fusion paradigm, data-level fusion, feature-level fusion, and decision-level fusion are the main directions explored in related studies (82). The framework of the third-generation diagnostic method for depression is illustrated in Figure 3.

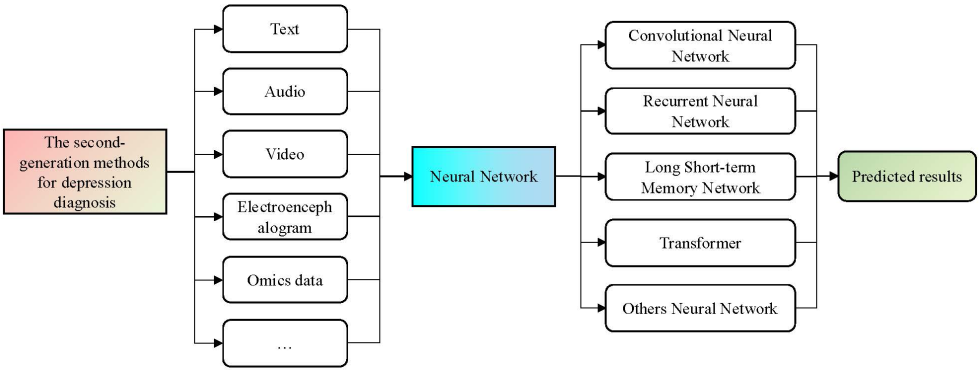


FIGURE 2
The second-generation framework for depression diagnosis methods. This generation of methods typically utilize the trained DL models on large-scale datasets as tools to represent different modal features. By using the end-to-end design and model training of neural networks, the diagnosis process of depression can be realized automatically.

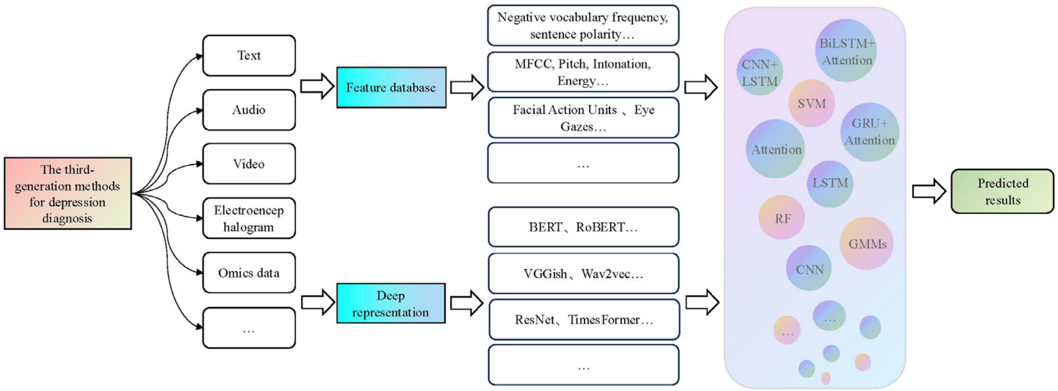


FIGURE 3
The third-generation framework for depression diagnosis methods. This generation of methods first combine prior features with depth features and then use a modal fusion strategy to integrate multiple model features as inputs for machine learning and neural network models used in the recognition of depression.

3.3.1 The data-level fusion methods

Data-level fusion, also known as early fusion, is an easily implementable method. In data-level fusion methods, different modalities of data are concatenated into integrated features using simple feature vectors, which are then fed into a neural network for depression recognition (83). The data-level fusion framework is shown in Figure 4.

In the research of data-level fusion, Zou et al. (84) extracted the 768-dimensional textual deep representation of interviewees based on interview content using the BERT model. They also extracted the 12-dimensional visual features and 25-dimensional audio features of interviewees using the Openface and eGemaps tools. Finally, they achieved automated identification of depression based on the Bi-LSTM model, with a recall rate of 0.89 and the F1-score of 0.88 on the Chinese Multimodal Depression Corpus (CMDC) dataset. Semwal et al. (85) collected over 10 million publicly available Twitter posts and approximately 10,000 images and obtained the feature vector representations of text and images using the Doc2Vec model and EfficientNetV2S architecture, respectively. The final accuracy of this approach reached 99.3%, with an F1 score of 95.61%. This was 17.36% higher than the best text model and 35.99% higher than the best image model, demonstrating that the information contained in text and images is complementary and can be effectively combined.

Although the data-level fusion method is relatively simple, it is a successful technique that combines features from different modalities for depression recognition. Compared to single-modality feature recognition methods, it considers more comprehensive information and enhances the reliability of diagnostic results. However, one drawback of data-level fusion is the high dimensionality of the combined feature vector. Considering this drawback, Joshi et al. (86) proposed data fusion and Principal Component Analysis (PCA) after data fusion, retaining 98% of the variance. The results showed that training a depression detection model on the dimensionality-reduced feature set can also improve system performance.

3.3.2 The feature-level fusion methods

Fusion at the feature-level, also known as intermediate fusion, combines the emotion features extracted from different modalities in a

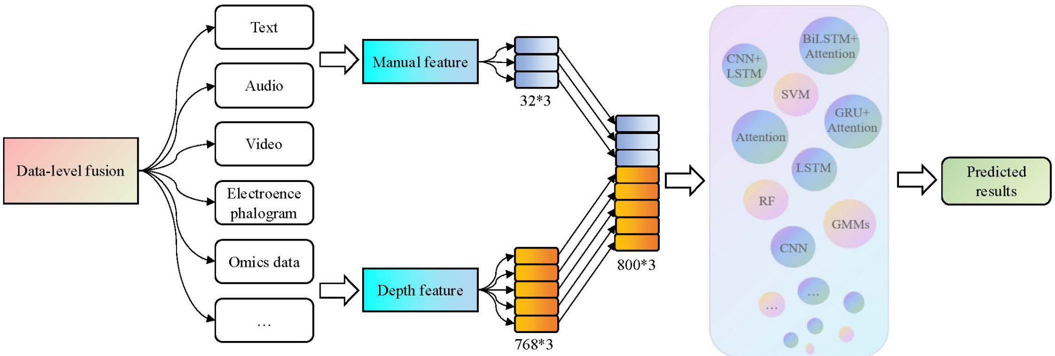


FIGURE 4
The data-level fusion process in the third-generation depression diagnosis methods. By using a simple feature vector concatenation method, varied features from different modalities and dimensions can be concentrated into comprehensive features and then input them into a neural network for depression recognition.

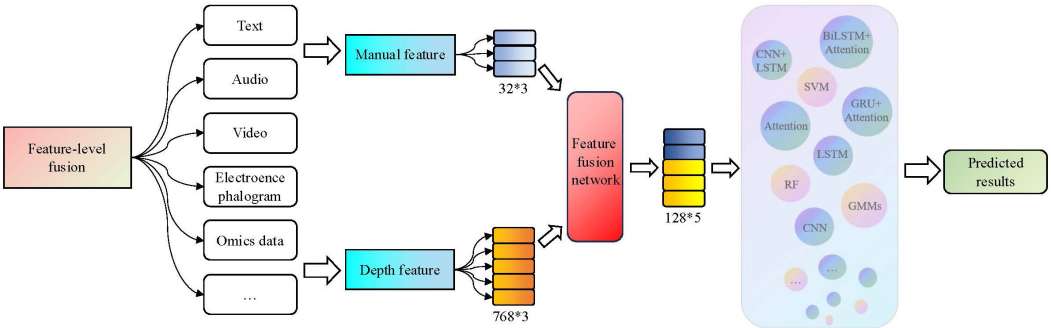


FIGURE 5
The feature-level fusion process in the third-generation depression diagnosis methods. By adding a feature fusion network, emotional features extracted from different modalities and dimensions are combined to form a comprehensive emotional feature.

certain way to form a comprehensive emotion feature as the input to identify networks. The fusion methods typically include neural networks, attention mechanisms, residual networks, and their combinations. The framework for feature-level fusion is illustrated in Figure 5.

In the research of feature-level fusion, Bucur et al. (87) proposed a method for identifying users with depression on social media platforms by combining text, image, and temporal information. They utilized EmoBERTa and Contrastive Language–Image Pre-training (CLIP) to encode the text and image, processed the encoded image and text sequences using a cross-modal encoder, and finally used a transformer encoder along with the relative posting time encoded into position embeddings to identify users with depression. Haque et al. (88) employed Causal Convolutional Neural Networks (C-CNN) to transform a variable-length sequence summarizing long-term auditory, visual, and linguistic elements into a fixed-size numerical vector, which was used to fuse the textual, audio, and 3D video features of facial landmark points. The model demonstrated promising performance on the DAIC-WOZ dataset.

Compared to the concatenation approach of feature vector fusion at the data-level, feature-level fusion utilizes intermediate fusion networks to transform and fuse different modality data, thus enhancing the correlation and complementary effects among the data. The effectiveness of the fusion network directly affects the diagnostic accuracy, but the specific fusion network required for different modal combinations or data may vary. Therefore, many researchers have begun investigating unified modality fusion encoders, such as the Visual-Image Encoder (89) and the Meta-Transformer (90) that fuses 12 modalities of data.

3.3.3 The decision-level fusion methods

Decision-level fusion is also known as late-fusion. Firstly, it constructs strong feature extraction models and classifiers for each modality and then selects appropriate fusion strategies to obtain the final results. Common fusion strategies include voting, weighting, and ML algorithms. The decision-level fusion framework is shown in Figure 6.

In terms of decision-level fusion research, Zhang et al. (91) proposed a deep forest based multimodal data fusion method for depression recognition to fuse multimodal data such as Histogram of Oriented Gradients (HOG), landmark, and AU of video frames. Zou et al. (84) integrated text content, acoustic features, and facial visual features, and implemented automatic alignment of multimodal features and depression diagnosis based on the Multiple Transformers (MulT) model. Yang et al. (82) proposed a multimodal fusion framework consisting of Deep Convolutional Neural Network (DCNN) and DNN models. For each mode,

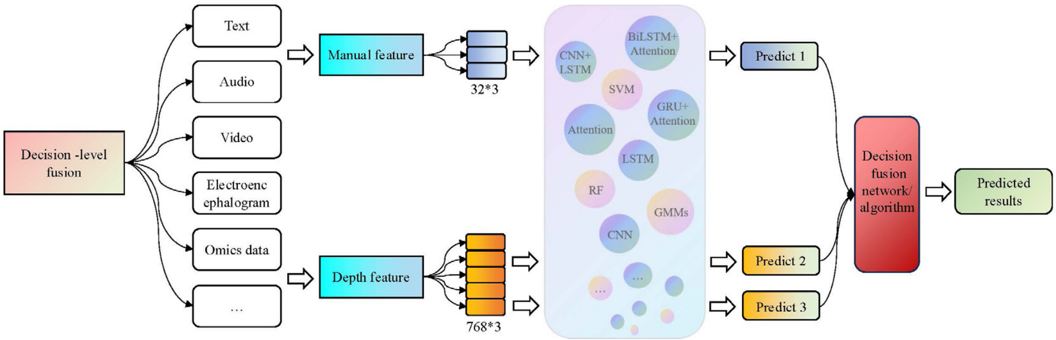


FIGURE 6
The decision-level fusion process in the third-generation depression diagnosis methods. It constructs strong feature extraction models and classifiers for each modality, and then selects appropriate fusion strategies to obtain the final results.

manually created feature descriptors are input into DCNN to learn advanced global features with compact dynamic information, and then the learned features are fed to DNN to predict the PHQ-8 score, with the final Mean Absolute Error (MAE) of the model of 5.163.

Decision-level fusion constructs a strong single modal feature extraction model based on the characteristics of each modality, which can effectively capture the internal information of each modality and help improve recognition rate. But one challenge is that decision-level fusion relies heavily on the selection of fusion algorithms, and the results are easily affected and interfered with by certain single modal information.

3.3.4 Other fusion methods

In addition to the methods of data-level fusion, feature-level fusion, and decision-level fusion, methods such as hybrid fusion, multi-level cascading, and a combination of multiple methods are currently also important research directions. For example, Wei et al. (92) introduced a sub attention model ConvBiLSTM that connects heterogeneous information based on the method of decision-level fusion. After incorporating this module, the diagnostic results of the model always outperform traditional decision-level fusion methods. Ray et al. (93) proposed a new multimodal depression prediction network based on multi-level attention networks, which integrates features of audio, video, and text modalities while learning correlations within and between modalities. The optimal performance was achieved in AVEC2019 (46), which was 17.52% higher than the baseline at that time.

3.3.5 Summary of the third-generation diagnostic methods for depression

The third-generation diagnostic methods combine the advantages of the previous two generations. The manually extracted features and DL model features have different attributes. Combining manual features with deep features can not only utilize previous experience but also mine hidden information in the original data. Secondly, the information of different modalities has complementarity, and through the strategy of modal fusion, feature information can be more comprehensively integrated to improve the accuracy of diagnostic results. However, the third-generation diagnostic methods still belong to the black box model and have weak interpretability. Although manual features are combined, their role in the diagnostic process cannot be explained. Secondly, while improving the accuracy of diagnosis, the strategy of feature combination and modal fusion also makes the entire diagnosis process more complex and far away from end-to-end goals.

3.4 Summary and analysis of diagnostic methods for the previous three generations

In view of the shortcomings of scale-based assessment, the method combined with AI technology has made great progress in diagnosis efficiency and objectivity, and a large number of experiments have also verified that the diagnosis effect of multimode is generally better than that of single mode, as shown in Table 2. From years of research in depression diagnosis and current clinical practices, language contents and facial expressions are commonly recognized as the modalities that are most closely related to depressive symptoms. In the future, AI models for depression diagnosis will integrate more types of data and knowledge, and thus interpret them more precisely. By incorporating various data sources that have been proven to correlate with depression, such as physiological signals and long-term behavioral information, AI models can more comprehensively capture the complex characteristics of depression.

Although AI-based depression diagnostic methods have improved diagnostic accuracy and objectivity, they still exhibit limitations or weaknesses in interpretability. This is primarily because classical machine learning algorithms and deep learning models have complex structures that involve multiple layers of non-linear transformations during data processing, making it difficult for humans to understand the decision-making process of the model. These models are often referred to as “black boxes” due to their lack of transparency in internal decision-making. This means that while these models can provide highly accurate predictions, their mechanisms and reasons for specific diagnoses are not easily understood. In clinical practice, doctors need to be accountable for diagnostic and treatment decisions. Interpretability ensures that doctors can clearly understand and explain each decision-making process. If the decision-making process of a model is opaque, doctors may question its reliability, and a lack of transparency can lead to unforeseen errors or biases, increasing the risk in clinical decision-making. Additionally, clinical depression diagnoses generally follow a broadly unified clinical diagnostic standard, which these models and methods often do not consider.

To address these issues, we combined the powerful and universal knowledge foundation of the LLMs with clinical diagnostic standards based on scale methods to propose the use of a new generation of depression auxiliary diagnosis method - “LLM as an AI agent for auxiliary diagnosis”. Our approach employs locally interpretable models in the decision-making process. By providing textual descriptions of intermediate outputs and decisions, doctors can verify the accuracy and reliability of the method. Moreover, by adhering to the diagnostic standards, we can improve the efficiency, accuracy, and interpretability of diagnosis.

4 Large language model as an AI agent for auxiliary diagnosis

4.1 The applications of LLM in the medical field

With the recent release of ChatGPT, some researchers have observed the powerful versatility and emergent potential of big language models, and have also begun to explore the possibility of introducing LLM into the medical field. Among them, in the medical question and answer tasks, Singhal et al. proposed Med-

TABLE 2 Comparison of results between single mode and multimodal fusion methods.

Method	Dataset	Measurement	Modality	Results
Anshul et al. (94)	Covid-19	Recall	T+S+D+U	0.78
			V+T+D+U	0.76
			V+T+S+U	0.84
			V+T+S+D	0.87
			V+S+D+U	0.89
			V+T+S+D+U	**0.91**
Zhang et al. (95)	Private	Recall	A	0.61
			V	0.64
			T	0.64
			A+V	**0.78**
Suri et al. (85)	Private	Accuracy	T	0.82
			I	0.63
			T+I	**0.99**
Bucur et al. (87)	Twitter (96)	Accuracy	T	0.87
			I	–
			T+I	**0.93**
Zou et al. (84)	CMDC	Recall	A	0.83
			V	0.71
			T	**0.90**
			A+V+T	0.89
Wei et al. (92)	DAIC-WOZ	Accuracy	A	0.77
			V	0.80
			T	–
			A+V+T	**0.83**
Ahmed et al. (97)		Recall	T+A	0.89
			T+E+F	0.92
			A+F	0.91
			T+A+E+F	**0.93**
Our proposed third-generation method	Private	Accuracy	A	0.64
			V	0.51
			T	0.73
			A+V+T	**0.78**

This table lists the major datasets used in AI research for diagnosing depression. The quality of the selected datasets was assessed based on their availability, representativeness of the studies, and relevance to the topic of this review. For example, the DAIC-WOZ dataset is a representative public dataset in the field of depression recognition. It has undergone ethical review, includes detailed annotations, and has been widely used in the AVEC series of competitions, demonstrating its reliability and broad applicability in this field. For private datasets, we ensured their usability by consulting the relevant references to verify the credibility of the data sources and the validity of the data analysis results. It is important to note that depression datasets generally exhibit some common issues, primarily inconsistencies in data collection methods and standards. These inconsistencies often arise from ethical and privacy concerns, leading to the inaccessibility of many datasets. Such issues are particularly prominent in the use of depression datasets and may affect the generalizability of research findings. Therefore, we have included this comparative explanation to help readers better understand the applicability and limitations of these datasets. T, Text; S, Emotional; E, Eyes; D, Depression Specific; U, User Specific; H, Head; A, Audio; V, Visual; I, Image; F, Landmark.

Bold indicates the maximum value of the indicator results in this experiment group.

Palm, which achieved an accuracy rate of 67.2% in the medical question and answer task (98). Subsequently, based on this model and some optimization strategies, Med-PaLM2 was proposed (99), which improved its accuracy by 19%, reaching an astonishing 86.5%. Under the tasks of Computed Tomography (CT) and X-ray image description, Microsoft's LLaVA-Med (100) can infer the pathological condition of patients based on the input image and generate questions and answers about the image. In terms of medical consultation, ChatDoctor (101) can understand medical terminology, procedures, and diagnostic situations. Patients can interact with the ChatDoctor model through chat interfaces and ask questions about health, symptoms, or medical conditions.

This type of LLM related to the medical field has undergone massive data training and fine-tuning, and can efficiently and accurately complete tasks such as medical consultation and medical image description. Meanwhile, due to the fact that the model output is a description of the content of the text, it greatly enhances the interpretability of the model output; It is also possible to verify the reliability of models or methods based on text output. Therefore, combining the automated diagnostic goals of depression with LLM will be a new direction for assisting diagnosis of depression.

4.2 AI agent for auxiliary diagnosis

The emergent abilities which integrated with the medical knowledge of the LLMs have shown us more possibilities for automated diagnosis of depression. But unlike simple medical knowledge Q&A tasks, the diagnosis of depression is the result of a comprehensive judgment of multiple aspects of information, including visual, language content, speech, and other behavioral characteristics, as well as external and internal factors such as personal condition, family situation, and indicator data. In addition, the diagnosis of depression also needs to follow medical clinical diagnostic standards in order to accurately identify the user's depression status. Therefore, the main purpose of depression diagnosis research is to construct a depression diagnosis method or system which integrates multi-source heterogeneous information, follows clinical diagnostic standards, and has interpretability. Using the LLM as a hub could effectively meet the above requirements.

In the AI agent method, we propose to use different foundation models to extract modal features and provide textual descriptions to enhance feature extraction ability and interpretability. This method also combines the powerful text understanding ability of the LLMs, integrates multi-source heterogeneous information, and achieves objective diagnosis from multiple dimensions. During the scale evaluation process, diagnostic criteria have been introduced to align with practical clinical diagnostic objectives. The proposed framework for the fourth-generation depression assistance and diagnostic system is shown in Figure 7.

In this system, we first combine the diagnostic process with the advantages of clinical scale evaluation methods through the interactions of the virtual doctors and interviewees, and finally evaluate the scale through LLM. In the practical research process of the method, we constructed a dialogue transcription model, two foundation feature extraction models (MDD audio model and MDD video model), to describe speech and visual characteristics. We also constructed an LLM-based ChatMDD (Chat Major Depressive Disorder) model for the diagnosis of depression and summarizing personal information. Based on the evaluation process and diagnostic criteria of the scale, the goal of automated diagnosis of depression is achieved. Implementing end-to-end interaction outside the entire system can improve diagnostic efficiency and reduce costs. By integrating multi-source heterogeneous information and the interaction of multiple large-scale models within the system, the situation of respondents can be analyzed from multiple dimensions. Besides, clinical diagnostic criteria were introduced to provide more accurate diagnostic support. The diagnostic results and all details are intuitively expressed in textual form, making the entire system highly interpretable and thus convincing the doctors and respondents with the validity of the

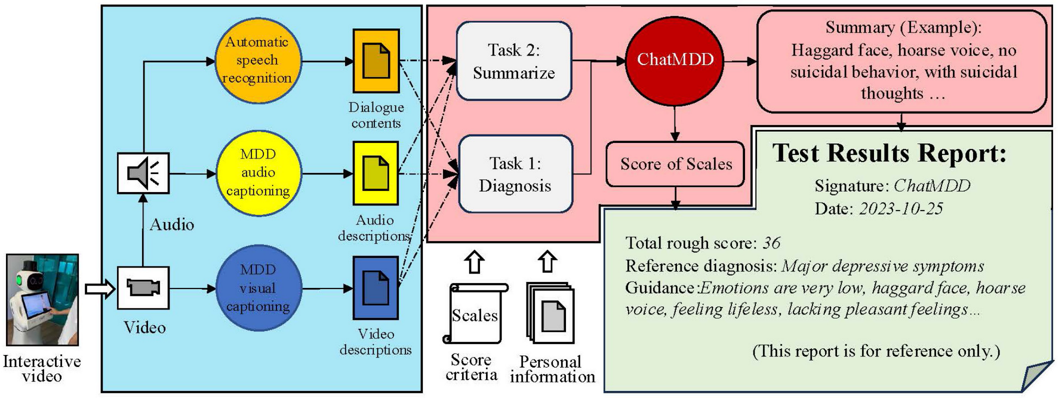

Automatic speech recognition
Dialogue contents
MDD audio captioning
Audio descriptions
MDD visual captioning
Video descriptions
Audio
Video
Interactive video
Task 2: Summarize
Task 1: Diagnosis
ChatMDD
Summary (Example): Haggard face, hoarse voice, no suicidal behavior, with suicidal thoughts …
Score of Scales
Test Results Report:
Signature: ChatMDD
Date: 2023-10-25
Total rough score: 36
Reference diagnosis: Major depressive symptoms
Guidance:Emotions are very low, haggard face, hoarse voice, feeling lifeless, lacking pleasant feelings...
(This report is for reference only.)
Scales
Score criteria
Personal information

FIGURE 7
The proposed framework for the fourth-generation depression assistance and diagnostic system. The entire system framework is divided into three parts. The first part involves extracting features from language, speech, and visual modalities using different basic models. The second part involves constructing cue templates based on features and diagnostic criteria through textual descriptions, followed by reasoning through the ChatMDD model. The third part involves formatting the results report.

diagnostic results. The methodology using the ChatMDD model is compared with other models and previous approaches reported in Table 3.

4.2.1 Feature extraction

The process of evaluating the scale was implemented in the form of muti-turns conversations with a virtual visual agent, during which we collected interview videos. After the MDD oriented conversion, the audio corresponding to the scale question can also be separated. By inputting the audio and video data into the foundation models of different modalities, the text descriptions of different modalities can be obtained. The text description results of the foundation model implementation features are shown in Figure 8.

For feature extraction methods and fusion techniques of different modalities, our approach diverges from previous approaches that represented features numerically. Instead, we adopted new audio captioner and visual captioner models to describe individual features textually. These textual descriptions were then integrated using a prompt template, which served as the input for LLMs. This method offers superior interpretability and scalability compared to the earlier generations of fusion techniques.

Specifically, the dialogue transcription model can more accurately transcribe complex professional terminology and rare disease types. The MDD audio model can analyze the audio information of the interviewee, capture the audio characteristics related to depression, and express it in the form of text, such as the interviewee's voice sounding somewhat depressed and low, the volume is low, and the speech is ambiguous. Consistent with the design concept of the language model, the MDD visual model can capture the visual characteristics of respondents, such as common visual features such as dullness, depression, desolation, and eye avoidance, as well as other features that cannot be intuitively detected such as eye movement changes, head posture, and movement unit changes, and express them in text. ChatMDD, through training on massive data and fine-tuning on depression datasets, can not only capture the language content information of interviewees, but also integrate other heterogeneous information such as personal information, medical knowledge, clinical diagnostic standards, audio features, visual features, etc., to assist in depression assessment, summarize interviewee information, and provide clinical guidance for reference. In addition, this architecture can be easily extended to other modalities in subsequent researches, such as electrophysiological foundation models, EEG foundation models, etc.

4.2.2 Prompt template

Integrating multi-source heterogeneous information as input to the ChatMDD also requires constructing appropriate prompt templates. In the above system, we constructed two different prompt templates to summarize the personal information of the respondents and use for depression assessment. In the summarized prompt template, we provide a task description and relevant information, allowing the ChatMDD model to automatically analyze the content and ultimately generate task results. In the diagnosis prompt template, we prompt the model to assume the role of a "depression diagnosis agent" to stimulate its professional capabilities. Additionally, we introduce diagnostic criteria to enhance the reliability of the model's assessments. The content of the Prompt template is shown in Figure 9. It should be noted that the content marked with colors is not complete, and the above is only for the convenience of understanding the construction method of the prompt template. In practical applications, suitable prompt templates can be constructed based on the data information needed.

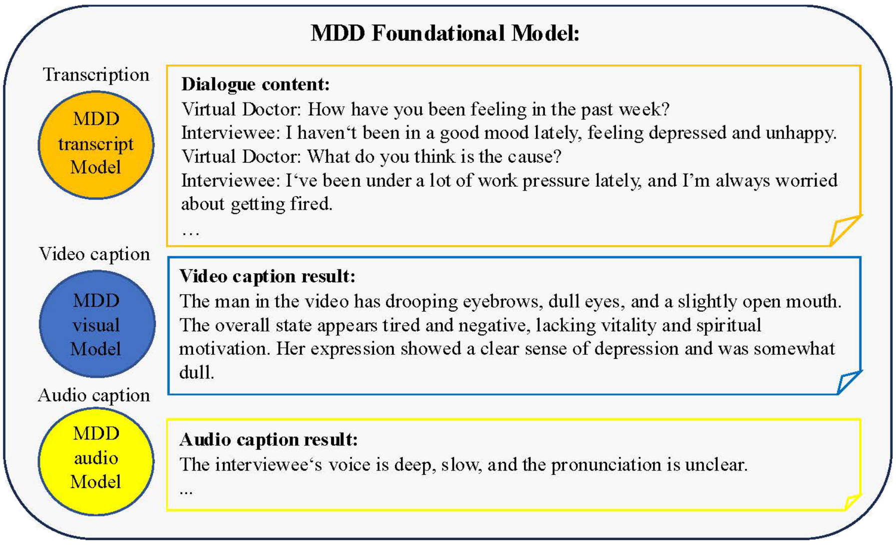
4.2.3 Diagnosis method and the report of ChatMDD

The diagnostic approach of ChatMDD comprises two methods. The first method involves analyzing all scale indicators at once,

TABLE 3 Comparison of ChatMDD with other models.

Method	Technology Base	Feature Processing	Objectivity	Diagnostic Standard	Interpretability	Scalability
Scale-based evaluation	Scale evaluation & expertise	Standardized questionnaires	Weak	Yes	Yes	No
The first-generation diagnostic methods	Classical machine learning algorithms	Manual feature extraction	Medium	No	Medium	No
The second-generation diagnostic methods	Deep learning models	Automated feature extraction	Medium	No	No	No
The third-generation diagnostic methods	ML + DL advantages	Mixed model for feature extraction	Medium	No	Weak	No
ChatMDD for auxiliary diagnosis	LLMs	Multi-modal data processing and analysis	Yes	Yes	Yes	Yes

Technology Base refers to the core technology underlying each method, such as scale-based evaluation or machine learning algorithms. Feature Processing describes how each method handles data, ranging from manual to automated processing. Objectivity is standardized as "No/Weak/Medium/Yes" to reflect the reliance on automated processing versus subjective judgment. Diagnostic Standard is simplified to "Yes/No" indicating whether the method has a clear diagnostic standard. Interpretability uses "Yes/Weak/No" to show how easily the model's outputs can be understood. Finally, Scalability is presented as "Yes/No" indicating the method's adaptability to varying data sizes and complexity levels. Additionally, the table does not suggest that LLM-based methods will replace all previous approaches; each method has its own suitable application scenarios. Scale-based evaluation is appropriate for traditional clinical assessments; methods based on classical machine learning algorithms are suitable for small to medium-scale data analysis and semi-automated processing; deep learning models are ideal for large-scale data analysis and automation; the third-generation hybrid models are well-suited for complex multi-source data analysis; and using LLMs as auxiliary diagnostic agents is most effective in high-demand clinical settings requiring high interpretability.



Dialogue content:
Virtual Doctor: How have you been feeling in the past week?
Interviewee: I haven't been in a good mood lately, feeling depressed and unhappy.
Virtual Doctor: What do you think is the cause?
Interviewee: I've been under a lot of work pressure lately, and I'm always worried about getting fired.
…

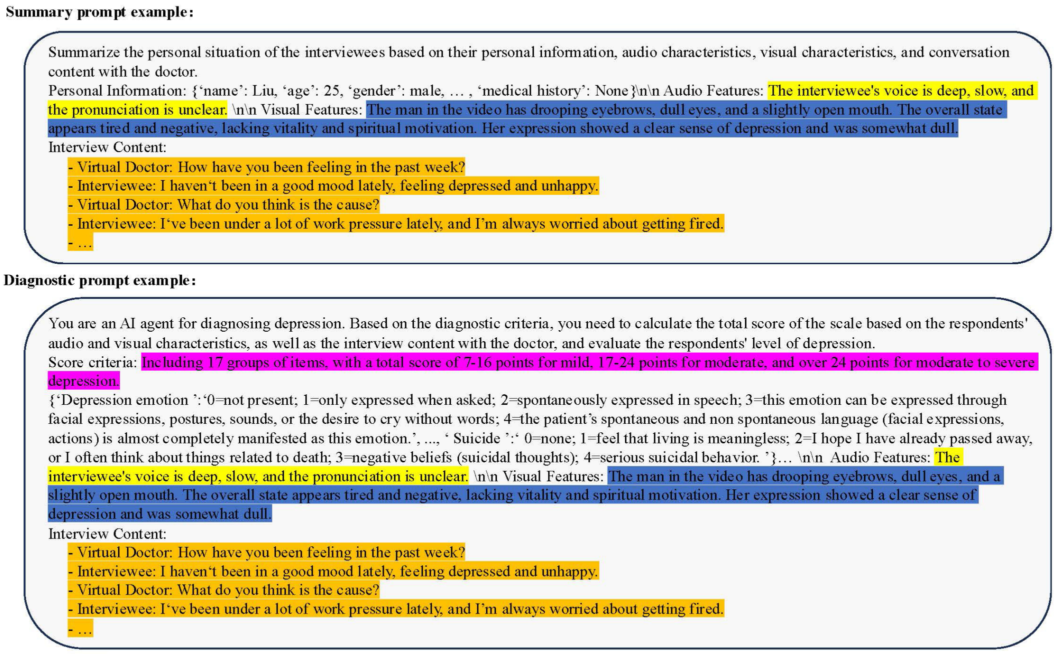
Video caption result:
The man in the video has drooping eyebrows, dull eyes, and a slightly open mouth. The overall state appears tired and negative, lacking vitality and spiritual motivation. Her expression showed a clear sense of depression and was somewhat dull.

Audio caption result:
The interviewee's voice is deep, slow, and the pronunciation is unclear.
...

FIGURE 8
Example illustration of the feature texturization results for the foundational model of MDD. The dialogue content includes interviews between virtual doctors and interviewees, and the questions involved are designed based on diagnostic criteria. The result of audio and video caption is a textual description of speech and visual features.

suitable for scales with fewer indicators. In this method, the inference process yields the final result in a single step, but the drawback is that the model requires high ability. The second method integrates all information and cyclically evaluates each indicator. Compared to the one-time approach, this method may yield more precise diagnostic results but requires more time for assessment.

In our system, the final diagnostic result obtained through the above process is shown in Figure 10. The ChatMDD model was used to combine all information, and each score was obtained based on the diagnostic criteria. The final score and diagnostic result were obtained. And for each score, we also provide criteria and criteria for judgment to enhance the interpretability of the model output, as shown in Figure 11.



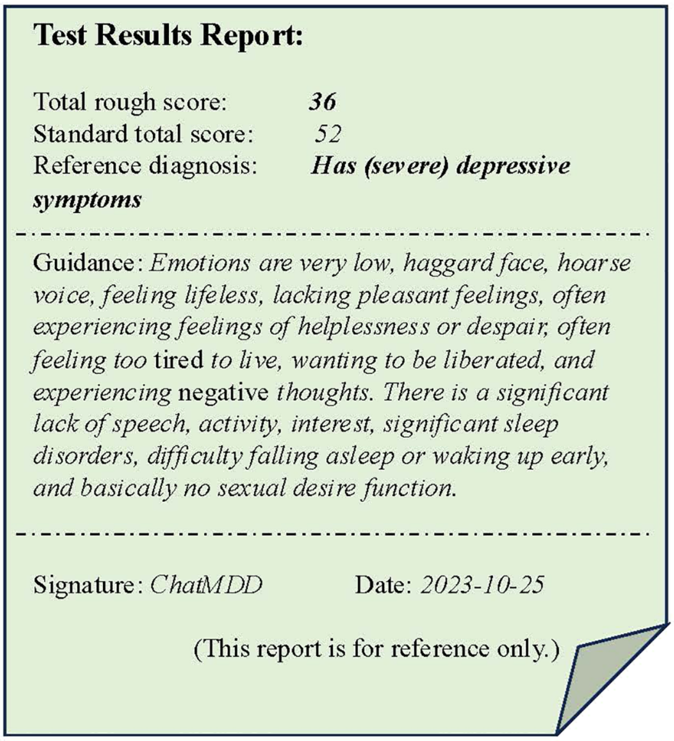
Summary prompt example:

Summarize the personal situation of the interviewees based on their personal information, audio characteristics, visual characteristics, and conversation content with the doctor.
Personal Information: {'name': Liu, 'age': 25, 'gender': male, … , 'medical history': None}\n\n Audio Features: The interviewee's voice is deep, slow, and the pronunciation is unclear. \n\n Visual Features: The man in the video has drooping eyebrows, dull eyes, and a slightly open mouth. The overall state appears tired and negative, lacking vitality and spiritual motivation. Her expression showed a clear sense of depression and was somewhat dull.
Interview Content:
- Virtual Doctor: How have you been feeling in the past week?
- Interviewee: I haven't been in a good mood lately, feeling depressed and unhappy.
- Virtual Doctor: What do you think is the cause?
- Interviewee: I've been under a lot of work pressure lately, and I'm always worried about getting fired.
- …

Diagnostic prompt example:

You are an AI agent for diagnosing depression. Based on the diagnostic criteria, you need to calculate the total score of the scale based on the respondents' audio and visual characteristics, as well as the interview content with the doctor, and evaluate the respondents' level of depression.
Score criteria: Including 17 groups of items, with a total score of 7-16 points for mild, 17-24 points for moderate, and over 24 points for moderate to severe depression.
{'Depression emotion ':'0=not present; 1=only expressed when asked; 2=spontaneously expressed in speech; 3=this emotion can be expressed through facial expressions, postures, sounds, or the desire to cry without words; 4=the patient's spontaneous and non spontaneous language (facial expressions, actions) is almost completely manifested as this emotion.', ..., ' Suicide ':' 0=none; 1=feel that living is meaningless; 2=I hope I have already passed away, or I often think about things related to death; 3=negative beliefs (suicidal thoughts); 4=serious suicidal behavior. '}… \n\n Audio Features: The interviewee's voice is deep, slow, and the pronunciation is unclear. \n\n Visual Features: The man in the video has drooping eyebrows, dull eyes, and a slightly open mouth. The overall state appears tired and negative, lacking vitality and spiritual motivation. Her expression showed a clear sense of depression and was somewhat dull.
Interview Content:
- Virtual Doctor: How have you been feeling in the past week?
- Interviewee: I haven't been in a good mood lately, feeling depressed and unhappy.
- Virtual Doctor: What do you think is the cause?
- Interviewee: I've been under a lot of work pressure lately, and I'm always worried about getting fired.
- …

FIGURE 9
Prompt Template Example: Summary Prompt Template (above), Diagnostic Prompt Template (below). The summary prompt template consists of four parts: personal basic information, audio features, visual features, and interview content. The diagnostic prompt template has changed personal basic information to diagnostic criteria and introduced role and task descriptions.



Test Results Report:

Total rough score: ***36***
Standard total score: *52*
Reference diagnosis: ***Has (severe) depressive symptoms***

Guidance: *Emotions are very low, haggard face, hoarse voice, feeling lifeless, lacking pleasant feelings, often experiencing feelings of helplessness or despair, often feeling too* tired *to live, wanting to be liberated, and experiencing* negative *thoughts. There is a significant lack of speech, activity, interest, significant sleep disorders, difficulty falling asleep or waking up early, and basically no sexual desire function.*

Signature: *ChatMDD* Date: *2023-10-25*

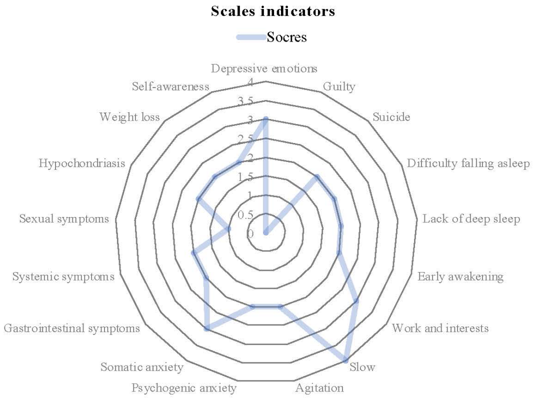
(This report is for reference only.)

FIGURE 10
A example of the test results reports. The test result report includes the total score and diagnostic results of the evaluation, as well as a summary of personal situations and some reference guidance.

4.2.4 Ethics, limitations, and challenges

When deploying AI models for diagnosing mental health conditions such as depression, several key ethical issues must be considered. Firstly, privacy and data protection are crucial given the sensitivity of mental health data. It is essential to ensure data anonymization to prevent patient identity disclosure. Secondly, informed consent is required; patients should be fully aware of how their data will be used and the associated risks. Additionally, the AI model's working principles, decision bases, and limitations must be transparently communicated to both patients and healthcare professionals.

However, it is important to acknowledge that there are still some limitations to this approach that need further improvement. For instance, due to the variable clinical manifestations of depression, LLMs require extensive and diverse datasets for training and fine-tuning to ensure accurate identification of a



Depression emotions: In the interview conversation, the interviewee mentioned that they were in a bad mood and showed a clear melancholic and somewhat dull visual characteristics. According to the score criteria for 'depressive emotions', the score is 3.
Score criteria: 0=not present; 1=only expressed when asked; 2=spontaneously expressed in speech; 3=this emotion can be expressed through facial expressions, postures, sounds, or the desire to cry without words; 4=the patient's spontaneous and non spontaneous language (facial expressions, actions) is almost completely manifested as this emotion.

Guilty: The interviewee did not mention anything about 'guilt'. According to the scoring criteria, the score is 0.
Score criteria: 0=not present; 1=blaming oneself, feeling that one has implicated others [including feeling disappointed in others]; 2=Believing that one has committed a crime or repeatedly reflecting on past mistakes and mistakes; 3=Believing that the current illness is a punishment for one's own mistakes, or having delusions of guilt; 4=delusions of guilt accompanied by accusations or threatening hallucinations.
…

FIGURE 11
Radar chart of scale indicators' results (left), along with the assessment outcomes and diagnostic criteria for each indicator (right).

wide range of symptoms. Additionally, challenges associated with dataset quality and bias handling, such as demographic imbalances, should be addressed to improve model generalizability and performance. These issues, along with a comprehensive comparison of ChatMDD and other state-of-the-art models in terms of accuracy, explainability, and clinical applicability, will be explored in future studies. Furthermore, clinical validation is the biggest barrier to the practical implementation of AI models in real-world clinical settings. Overcoming this challenge would require large-scale clinical trials and validation, which is something we have been actively working on. Only through thorough clinical validation can we ensure the effectiveness and reproducibility of these models in real-world scenarios, thereby facilitating their widespread adoption in clinical practice.

5 Summary

In this paper, we first introduced a scale-based depression diagnosis method, systematically reviewed the research of AI technology in depression diagnosis, and pointed out that the comprehensive, objective, and interpretable results are the problems that need to be addressed in the current depression diagnosis research. However, with the development of LLM technology, many issues identified in previous studies could be addressed by applying it to depression diagnosis. Finally, we propose a framework for the fourth-generation depression auxiliary diagnosis system based on the ChatMDD model and make some further attempts in this area. The new method of AI agent has enormous potentials, but it also faces the following two challenges: (1) high-quality training data is critical for improving the performance of LLM, and the collection of depression patient data is a very difficult and time-consuming task; and (2) the current method of AI agent is to use LLM to process the text data converted from the foundation model. Therefore, by constructing a multimodal LLM as the agent, it will be more in line with the idea of multi-dimensional comprehensive diagnosis.

Author contributions

HL: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal Analysis, Data curation, Conceptualization. HW: Investigation, Writing – review & editing, Validation, Project administration, Methodology, Conceptualization. ZY: Writing – review & editing, Validation, Investigation, Data curation. ZR: Writing – original draft, Validation, Investigation, Data curation. YD: Writing – review & editing, Resources, Data curation, Investigation. GZ: Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization. MDL: Writing – review & editing, Visualization, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported in part by the fund from the China Precision Medicine Initiative (2016YFC0906300), the National Major Talent Engineering Expert Workstation of Taiyuan Municipality government, and the State Key Laboratory for Diagnosis and Treatment of Infectious Diseases of the First Affiliated Hospital of Zhejiang University.

Acknowledgments

We thank the Qwen Team of Alibaba Group in providing the Qwen2-VL (v. 7B-Instruct) for both visual and audio feature extraction, and the Large Language Model Qwen2 (v. 7B) for Inference estimation. We also thank the Speech Lab of Alibaba Group in providing the Paraformer (v.speech_seaco_paraformer_large_asr_nat-zh-cn-16k-common-vocab8404-pytroch) for speech transcription.

Conflict of interest

Authors HL and YD were employed by Shanxi Tongchuang Technology Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Beck AT, Alford BA. *Depression: causes and treatment*. University of Pennsylvania Press (2009).

2. Kim JW, Suzuki K, Kavalali ET, Monteggia LM. Ketamine: mechanisms and relevance to treatment of depression. *Annu Rev Med.* (2024) 75:129–43. doi: 10.1146/annurev-med-051322-120608

3. Who. *Depression and other common mental disorders: global health estimates*. Geneva, Switzerland: WHO Document Production Services (2017).

4. Hasin DS, Sarvet AL, Meyers JL, Saha TD, Ruan WJ, Stohl M, et al. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry.* (2018) 75:336–46. doi: 10.1001/jamapsychiatry.2017.4602

5. Huang Y, Wang Y, Wang H, Liu Z, Yu X, Yan J, et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study. *Lancet Psychiatry*. (2019) 6:211–24. doi: 10.1016/S2215-0366(18)30511-X

6. Strawbridge R, Mccrone P, Ulrichsen A, Zahn R, Eberhard J, Wasserman D, et al. Care pathways for people with major depressive disorder: a European Brain Council Value of Treatment study. *Eur Psychiatry*. (2022) 65:1–21. doi: 10.1192/j.eurpsy.2022.28

7. Zulfiker MS, Kabir N, Biswas AA, Nazneen T, Uddin MS. An in-depth analysis of machine learning approaches to predict depression. *Curr Res Behav Sci*. (2021) 2:100044. doi: 10.1016/j.crbeha.2021.100044

8. Eder J, Pfeiffer L, Wichert SP, Keeser B, Simon MS, Popovic D, et al. Deconstructing depression by machine learning: the POKAL-PSY study. *Eur Arch Psychiatry Clin Neurosci*. (2024) 274:1153–65. doi: 10.1007/s00406-023-01720-9

9. Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. *J Affect Disord*. (2021) 281:618–22. doi: 10.1016/j.jad.2020.11.104

10. Benrimoh D, Kleinerman A, Furukawa TA, Reynolds Iii CF, Lenze EJ, Karp J, et al. Towards outcome-driven patient subgroups: a machine learning analysis across six depression treatment studies. *Am J Geriatric Psychiatry*. (2024) 32:280–92. doi: 10.1016/j.jagp.2023.09.009

11. Gutiérrez R, Rampérez V, Paggi H, Lara JA, Soriano J. On the use of information fusion techniques to improve information quality: Taxonomy, opportunities and challenges. *Inf Fusion*. (2022) 78:102–37. doi: 10.1016/j.inffus.2021.09.017

12. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. (2020) 33:1877–901. doi: 10.48550/arXiv.2005.14165

13. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. (2023). doi: 10.48550/arXiv.2303.08774

14. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. (2023). doi: 10.48550/arXiv.2307.09288

15. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. (2024). doi: 10.48550/arXiv.2407.21783

16. Du Z, Qian Y, Liu X, Ding M, Qiu J, Yang Z, et al. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*. (2021). doi: 10.48550/arXiv.2103.10360

17. Glm T, Zeng A, Xu B, Wang B, Zhang C, Yin D, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793* (2024). doi: 10.48550/arXiv.2406.12793

18. Dawson-Squibb J-J, Davids EL, Viljoen M, Rice K, Stein DJ. The WHO international classification of diseases 11th revision (ICD-11). In: *Handbook of clinical child psychology: Integrating theory and research into practice*. Cham: Springer (2023). p. 53–78.

19. Regier DA, Narrow WE, Kuhl EA, Kupfer DJ. The conceptual development of DSM-V. *Am J Psychiatry*. (2009) 166:645–50. doi: 10.1176/appi.ajp.2009.09020279

20. Taboada Gjorup AL, Tolentino Junior JC, Van Duinkerken E, Marques AC, Do Carmo Filho A, Joaquim AM, et al. Association between attention performance and the different dimensions of DSM-5 depression symptoms. *Front Psychiatry*. (2023) 14:1291670. doi: 10.3389/fpsyt.2023.1291670

21. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry*. (1998) 59:22–33. doi: 10.4088/JCP.v59n0105

22. Widyowati A, Peristiowati Y, Ningtiyas LK. Pengaruh terapi konseling terhadap tingkat depresi pasien melalui pemeriksaan mini international neuropsychiatric interview. *J Telenursing (JOTING)*. (2024) 6:429–37. doi: 10.31539/joting.v6i1.9096

23. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. (2001) 16:606–13. doi: 10.1046/j.1525-1497.2001.016009606.x

24. Palmer EOC, Ker S, Renteria ME, Carmody T, Rush AJ. Psychometric evaluation and linking of the PHQ-9, QIDS-C, and VQIDS-C in a real-world population with major depressive disorder. *Neuropsychiatr Dis Treat*. (2024) 20:671–87. doi: 10.2147/NDT.S444223

25. Olatunji BO, Deacon BJ, Abramowitz JS, Tolin DF. Dimensionality of somatic complaints: factor structure and psychometric properties of the Self-Rating Anxiety Scale. *J Anxiety Disord*. (2006) 20:543–61. doi: 10.1016/j.janxdis.2005.08.002

26. Deng J. Effects of anxiety and depression on positioning error of radiotherapy. *Int J Front Med*. (2024) 6:56–61. doi: 10.25236/IJFM.2024.060309

27. Zung WW, Richards CB, Short MJ. Self-rating depression scale in an outpatient clinic: further validation of the SDS. *Arch Gen Psychiatry*. (1965) 13:508–15. doi: 10.1001/archpsyc.1965.01730060026004

28. Sepehry AA. Self-rating depression scale (SDS). In: *Encyclopedia of quality of life and well-being research*. Cham: Springer (2024). p. 6269–76.

29. Carrozzino D, Patierno C, Fava GA, Guidi J. The hamilton rating scales for depression: A critical review of clinimetric properties of different versions. *Psychother Psychosom*. (2020) 89:133–50. doi: 10.1159/000506879

30. Wu Y, Levis B, Sun Y, He C, Krishnan A, Neupane D, et al. Accuracy of the Hospital Anxiety and Depression Scale Depression subscale (HADS-D) to screen for major depression: systematic review and individual participant data meta-analysis. *BMJ*. (2021) 373:n972. doi: 10.1136/bmj.n972

31. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. (2015) 349:255–60. doi: 10.1126/science.aaa8415

32. Fan H, Zhang X, Xu Y, Fang J, Zhang S, Zhao X, et al. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Information Fusion*. (2024) 104:102161. doi: 10.1016/j.inffus.2023.102161

33. Gao S, Calhoun VD, Sui J. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neurosci Ther*. (2018) 24:1037–52. doi: 10.1111/cns.2018.24.issue-11

34. D'amico A, Capua MD, Nardo ED, Rosak-Szyrocka J, Tortorella G, Festa G, et al. Competitive advantage in healthcare based on augmentation of clinical images with artificial intelligence: case study of the'Sambias' project. *International J Managerial Financial Accounting*. (2024) 16:1–16. doi: 10.1504/IJMFA.2024.135364

35. Trifu RN, Nemeş B, Bodea-Haţegan C, Cozman D. Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *J Evidence-Based Psychotherapies*. (2017) 17:105–28. doi: 10.24193/jebp.2017.1.7

36. Yin C, Zhang S, Zhu G, Zhang B. A method of mining sentiment word for depression patients based on word frequency-polarity intensity. *J Taiyuan Univ Technol*. (2021) 52:105–10. doi: 10.16355/j.cnki.issn1007-9432tyut.2021.01.014

37. Liu H, Zhou G, Qiao J, Zhang G. Chinese depression text classification based on domain emotion dictionary and word feature fusion. *J North Univ China (Natural Sci Edition)*. (2022) 43:522–9. doi: 10.3969/j.issn.1673-3193.2022.06.007

38. He K, Guan Y, Gong R. A chinese text classification algorithm based on weight preprocessing. *Comput Technol And Dev*. (2022) 32:40–45+53. doi: 10.3969/j.issn.1673-629X.2022.03.007

39. Zhang C, Chen X, Wang S, Hu J, Wang C, Liu X. Using CatBoost algorithm to identify middle-aged and elderly depression, national health and nutrition examination survey 2011–2018. *Psychiatry Res*. (2021) 306:114261. doi: 10.1016/j.psychres.2021.114261

40. Diao Y, Wang H, Dong J, Zhu Y, Shao Q, Feng L, et al. Advances in the application of machine learning in auxiliary diagnosis of depression. *Chin J Med Phys*. (2022) 39:257–64. doi: 10.3969/j.issn.1005-202X.2022.02.021

41. Zhang L, Duvvuri R, Chandra K, Nguyen T, Ghomi R. Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. *SSRN Electronic J*. (2019) 37:657–69. doi: 10.2139/ssrn.3408093

42. Williamson JR, Godoy E, Cha M, Schwarzentruber A, Khorrami P, Gwon Y, et al. Detecting depression using vocal, facial and semantic communication cues, in: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. Amsterdam, The Netherlands: Association for Computing Machinery. (2016).

43. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*. (2015) 71:10–49. doi: 10.1016/j.specom.2015.03.004

44. Shankayi R, Vali M, Salimi M, Malekshahi M. Identifying depressed from healthy cases using speech processing, in: *2012 19th Iranian Conference of Biomedical Engineering (ICBME)*. Tehran, Iran: IEEE. (2012) pp. 242–5.

45. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans Affect Computing*. (2016) 7:190–202. doi: 10.1109/TAFFC.2015.2457417

46. Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition, in: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. Nice, France: Association for Computing Machinery. (2019).

47. Sumali B, Mitsukura Y, Liang K-C, Yoshimura M, Kitazawa M, Takamiya A, et al. Speech quality feature analysis for classification of depression and dementia patients. *Sensors*. (2020) 20:3599. doi: 10.3390/s20123599

48. Cummins N, Epps J, Breakspear M, Goecke R. An investigation of depressed speech detection: features and normalization, in: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*. Florence, Italy: International Speech Communication Association. (2011).

49. Jiang H, Hu B, Liu Z, Wang G, Zhang L, Li X, et al. Detecting depression using an ensemble logistic regression model based on multiple speech features. *Comput Math Methods Med*. (2018) 2018:6508319. doi: 10.1155/2018/6508319

50. Carvalho N, Laurent E, Noiret N, Chopard G, Haffen E, Bennabi D, et al. Eye movement in unipolar and bipolar depression: A systematic review of the literature. *Front Psychol*. (2015) 6:1809. doi: 10.3389/fpsyg.2015.01809

51. Scherer S, Stratou G, Mahmoud M, Boberg J, Gratch J, Rizzo A, et al. Automatic behavior descriptors for psychological disorder analysis, in: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Shanghai, China: IEEE. (2013) pp. 1–8.

52. Pampouchidou A, Simos PG, Marias K, Meriaudeau F, Yang F, Pediaditis M, et al. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Trans Affect Computing*. (2019) 10:445–70. doi: 10.1109/T-AFFC.5165369

53. Baltrusaitis T, Zadeh A, Lim YC, Morency LP. OpenFace 2.0: facial behavior analysis toolkit, in: *IEEE International Conference on Automatic Face & Gesture Recognition*. Xi'an, China: IEEE. (2018). pp. 59–66.

54. Wang Q, Yang H, Yu Y. Facial expression video analysis for depression detection in Chinese patients. *J Visual Communication Image Representation*. (2018) 57:228–33. doi: 10.1016/j.jvcir.2018.11.003

55. Li J, Liu Z, Ding Z, Wang G. A novel study for MDD detection through task-elicited facial cues, in: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Madrid, Spain: IEEE. (2021) pp. 1003–8.

56. Li X, Fan Q. Application progress of machine learning in the study of facial features of patients with depression. *J ShangHai Jiao Tong Univ (Medical Science)*. (2022) 42:124–9. doi: 10.3969/j.issn.1674-8115.2022.01.019

57. Zhao J, Zhou L, Wu M, Wang X, Meng X. AI-based assisted diagnostic methods for depression. *J Northwest Univ (Natural Sci Edition)*. (2023) 53:325–35. doi: 10.16152/j.cnki.xdxbzr.2023-03-002

58. Bhaumik R, Jenkins LM, Gowins JR, Jacobs RH, Barba A, Bhaumik DK, et al. Multivariate pattern analysis strategies in detection of remitted major depressive disorder using resting state functional connectivity. *NeuroImage: Clin*. (2017) 16:390–8. doi: 10.1016/j.nicl.2016.02.018

59. Wei M, Qin J, Yan R, Li H, Yao Z, Lu Q. Identifying major depressive disorder using Hurst exponent of resting-state brain networks. *Psychiatry Research: Neuroimaging*. (2013) 214:306–12. doi: 10.1016/j.pscychresns.2013.09.008

60. Akbari H, Sadiq MT, Ur Rehman A, Ghazvini M, Naqvi RA, Payan M, et al. Depression recognition based on the reconstruction of phase space of EEG signals and geometrical features. *Appl Acoustics*. (2021) 179:108078. doi: 10.1016/j.apacoust.2021.108078

61. Dipnall JF, Pasco JA, Berk M, Williams LJ, Dodd S, Jacka FN, et al. Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PLoS One*. (2016) 11:e0148195. doi: 10.1371/journal.pone.0148195

62. Yang L, Jiang D, He L, Pei E, Oveneke MC, Sahli H. Decision tree based depression classification from audio video and language information, in: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. Amsterdam, The Netherlands: Association for Computing Machinery. (2016).

63. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. *The distress analysis interview corpus of human and computer interviews*. Reykjavik, Iceland: LREC, European Language Resources Association (ELRA). (2014). p. 3123–8.

64. Alghowinem S, Goecke R, Wagner M, Epps J, Hyett M, Parker G, et al. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Trans Affect Computing*. (2018) 9:478–90. doi: 10.1109/T-AFFC.5165369

65. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans intelligent Syst Technol*. (2011) 2:1–27. doi: 10.1145/1961189.1961199

66. Zhao J, Su W, Jia J, Zhang C, Lu T. Research on depression detection algorithm combine acoustic rhythm with sparse face recognition. *Cluster Computing*. (2019) 22:7873–84. doi: 10.1007/s10586-017-1469-0

67. Xu J. *The Study of Depression Analysis Based on Deep Learning and Social Text*. D, Engineering at Harbin Institute of Technology (2022).

68. Ive J, Gkotsis G, Dutta R, Stewart R, Velupillai S. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health, in: *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. New Orleans, LA: Association for Computational Linguistics. (2018) pp. 69–77.

69. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. *BioMed Inform Insights*. (2018) 10:1178222618792860. doi: 10.1177/1178222618792860

70. Chlasta K, Wołk K, Krejtz I. Automated speech-based screening of depression using deep convolutional neural networks. *Proc Comput Sci*. (2019) 164:618–28. doi: 10.1016/j.procs.2019.12.228

71. Valstar M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres Torres M, et al. (2016). AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Association for Computing Machinery*. The Netherlands: Amsterdam. pp. 3–10.

72. Lu X, Shi D, Liu Y, Yuan J, Dong Q, Ma X. Speech depression tendency recognition based on attentional residual network. *J Chin Comput Syst*. (2022) 43:1602–8. doi: 10.20009/j.cnki.21-1106/TP.2021-0375

73. Li J, Fu X. Audio depression recognition based on deep learning. *Comput Appl Software*. (2019) 36:161–7. doi: 10.3969/j.issn.1000-386x.2019.09.029

74. Gao Y. *Depression Recognition Research Based on Multi-scale Facial Features*. D, Lanzhou University (2021).

75. Xu X. *Depression Recognition Algorithm Based on Facial Deep Spatio-Temporal Features*. D, Hebei University of Technology (2021).

76. De Melo WC, Granger E, Hadid A. Combining global and local convolutional 3d networks for detecting depression from facial expressions, in: *2019 14th ieee international conference on automatic face & gesture recognition*. Lille, France: IEEE. (2019). pp. 1–8.

77. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H, Puthankattil SD. Automated EEG-based screening of depression using deep convolutional neural network. *Comput Methods Programs Biomedicine*. (2018) 161:103–13. doi: 10.1016/j.cmpb.2018.04.012

78. Lin E, Kuo PH, Liu YL, Yu YW, Yang AC, Tsai SJ. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Front Psychiatry*. (2018) 9:290. doi: 10.3389/fpsyt.2018.00290

79. Wang Y, Wang Z, Li C, Zhang Y, Wang H. A multimodal feature fusion-based method for individual depression detection on sina weibo, in: *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*. Austin, TX, USA: IEEE. (2020) pp. 1–8.

80. Hao Y. *Depression Recognition Based on Multi-modularity*. D, Yangzhou University (2022).

81. Tao Y, Yang M, Li H, Wu Y, Hu B. DepMSTAT: multimodal spatio-temporal attentional transformer for depression detection. *IEEE Trans Knowledge Data Eng*. (2024) 36:2956–66. doi: 10.1109/TKDE.2024.3350071

82. Yang L, Jiang D, Xia X, Pei E, Oveneke MC, Sahli H. Multimodal measurement of depression using deep learning models, in: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. Mountain View, California, USA: Association for Computing Machinery. (2017).

83. Morales M, Scherer S, Levitan R. A linguistically-informed fusion approach for multimodal depression detection, in: *proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. New Orleans, LA: Association for Computational Linguistics. (2018) pp. 13–24.

84. Zou B, Han J, Wang Y, Liu R, Zhao S, Feng L, et al. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Trans Affect Computing*. (2022), 14:2823–38. doi: 10.1109/TAFFC.2022.3181210

85. Suri M, Chaudhary D, Gorton I, Kumar B. *Multimodal Analysis and Modality Fusion for Detection of Depression from Twitter Data*. Vol. 1–5. Association for the Advancement of Artificial Intelligence (2023).

86. Joshi J, Goecke R, Alghowinem S, Dhall A, Wagner M, Epps J, et al. Multimodal assistive technologies for depression diagnosis and monitoring. *J Multimodal User Interfaces*. (2013) 7:217–28. doi: 10.1007/s12193-013-0123-2

87. Bucur A-M, Cosma A, Rosso P, Dinu LP. It's just a matter of time: detecting depression with time-enriched multimodal transformers, in: *European Conference on Information Retrieval*. Dublin, Ireland: Springer-Verlag. (2023). pp. 200–15.

88. Haque A, Guo M, Miner AS, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv preprint arXiv:1811.08592*. (2018). doi: 10.48550/arXiv.1811.08592

89. Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision, in: *International Conference on Machine Learning: PMLR*, (No: PMLR). (2021) pp. 5583–94.

90. Zhang Y, Gong K, Zhang K, Li H, Qiao Y, Ouyang W, et al. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*. (2023). doi: 10.48550/arXiv:2307.10802

91. Zhang Y, Nie Y, Sun B, He J, Yang B. Decision level fusion depression evaluation method based on deep forest multimodal data. *J Beijing Normal Univ (Natural Science)*. (2018) 54:606–11. doi: 10.16360/j.cnki.jbnuns.2018.05.008

92. Wei P-C, Peng K, Roitberg A, Yang K, Zhang J, Stiefelhagen R. Multi-modal depression estimation based on sub-attentional fusion, in: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland. (2023). 623–39.

93. Ray A, Kumar S, Reddy R, Mukherjee P, Garg R. Multi-level attention network using text, audio and video for depression prediction, in: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. Nice, France: Association for Computing Machinery. (2019) pp. 81–88.

94. Anshul A, Pranav GS, Rehman MZU, Kumar N. A multimodal framework for depression detection during COVID-19 via harvesting social media. *IEEE Trans Comput Soc Syst*. (2024) 11:2872–88. doi: 10.1109/TCSS.2023.3309229

95. Zhang Z, Zhang S, Ni D, Wei Z, Yang K, Jin S, et al. Multimodal sensing for depression risk detection: integrating audio, video, and text data. *Sensors (Basel)*. (2024) 24:3714. doi: 10.3390/s24123714

96. Gui T, Zhu L, Zhang Q, Peng M, Zhou X, Ding K, et al. Cooperative multimodal approach to depression detection in twitter, in: *Proceedings of the AAAI conference on artificial intelligence*. Honolulu, Hawaii, USA: AAAI Press. (2019). pp. 110–7.

97. Ahmed S, Abu Yousuf M, Monowar MM, Hamid A, Alassafi MO. Taking all the factors we need: A multimodal depression classification with uncertainty approximation. *IEEE Access*. (2023) 11:99847–61. doi: 10.1109/ACCESS.2023.3315243

98. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*. (2022). doi: 10.48550/arXiv:2212.13138

99. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*. (2023). doi: 10.48550/arXiv:2305.09617

100. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv Neural Inf Process Syst*. (2024) 36:28541–64. doi: 10.48550/arXiv:2306.00890

101. Li Y, Li Z, Zhangi K, Dan R, Zhang Y. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*. (2023). doi: 10.48550/arXiv:2303.14070