

AI Models for Depressive Disorder Detection and Diagnosis: A Review

Dorsa Macky Aleagha^{1*}, Payam Zohari^{1†}, Mostafa Haghir Chehreghani^{1‡}

¹Department of Computer Engineering, Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran

Abstract

Major Depressive Disorder is one of the leading causes of disability worldwide, yet its diagnosis still depends largely on subjective clinical assessments. Integrating Artificial Intelligence (AI) holds promise for developing objective, scalable, and timely diagnostic tools. In this paper, we present a comprehensive survey of state-of-the-art AI methods for depression detection and diagnosis, based on a systematic review of 55 key studies. We introduce a novel hierarchical taxonomy that structures the field by primary clinical task (diagnosis vs. prediction), data modality (text, speech, neuroimaging, multimodal), and computational model class (e.g., graph neural networks, large language models, hybrid approaches). Our in-depth analysis reveals three major trends: the predominance of graph neural networks for modeling brain connectivity, the rise of large language models for linguistic and conversational data, and an emerging focus on multimodal fusion, explainability, and algorithmic fairness. Alongside methodological insights, we provide an overview of prominent public datasets and standard evaluation metrics as a practical guide for researchers. By synthesizing current advances and highlighting open challenges, this survey offers a comprehensive roadmap for future innovation in computational psychiatry.

keywords Depressive disorder, Artificial Intelligence (AI), multimodal analysis, Graph Neural Networks (GNNs), Large Language Models (LLMs), computational psychiatry

1 Introduction

Major Depressive Disorder (MDD), commonly known as depression, is a significant global health challenge characterized by a persistent low mood, a loss of interest in previously enjoyable activities (anhedonia), and a range of emotional, cognitive, and physical symptoms [1]. This debilitating condition negatively affects how individuals feel, think, and act, and can lead to severe impairments in daily functioning, including disturbances in sleep, appetite, and concentration [2]. As a leading cause of disability worldwide, depression contributes substantially to the overall global burden of disease, impacting not only personal well-being but also imposing a considerable economic strain

*Email: dorsa.macky@gmail.com

†Email: payam.zohari@aut.ac.ir

‡Corresponding author. Email: mostafa.chehreghani@aut.ac.ir

through reduced productivity and increased healthcare costs [3, 4]. Given its profound impact, the early and accurate detection of depressive disorders is crucial for initiating timely interventions, improving patient outcomes, and mitigating the broader societal consequences.

In recent years, the field of computational psychiatry has seen a surge of interest in leveraging Artificial Intelligence (AI) to address this challenge. The growing availability of large-scale digital datasets—spanning social media text, clinical interview transcripts, wearable sensor data, and neuroimaging—has created a fertile ground for the application of advanced machine learning models [5]. AI offers the promise of moving beyond traditional, subjective diagnostic methods by identifying complex, objective, and often subtle patterns that are indicative of depression. This has led to the rapid development of sophisticated models capable of analyzing diverse data modalities with increasing accuracy and efficiency [6].

Despite these advancements, the field faces significant challenges. Integrating heterogeneous, high-dimensional data from multiple sources remains a complex task, often requiring specialized fusion techniques to be effective [7]. Furthermore, issues of data privacy, demographic bias in algorithms, and the “black-box” nature of many deep learning models pose considerable hurdles to their clinical implementation, underscoring the pressing need for models that are not only accurate but also fair, transparent, and interpretable [5, 8].

A number of valuable surveys review the intersection of AI and depression detection. Some provide broad overviews of machine learning applications in mental health [6], while others focus on specific data modalities like social media [9] or neuroimaging [10]. However, to the best of our knowledge, no existing survey offers a comprehensive, hierarchical taxonomy that systematically organizes the field first by the clinical task (Diagnosis vs. Prediction), then by data modality, and finally by the class of computational model employed. This structural gap makes it difficult to navigate the landscape, compare methodologies, and identify specific research opportunities.

This survey aims to fill this critical gap by providing the first structured review of the field based on a novel hierarchical classification framework. Our paper makes several key contributions. First, we define and differentiate the primary tasks, data types, and methodological approaches prevalent in the literature. Second, we present a comprehensive, multi-level classification of 55 recent and major papers. Third, we conduct an in-depth review of these works, following our tree-based taxonomy to discuss the innovations and findings within each specific sub-domain. Finally, we introduce the most commonly used public datasets and evaluation metrics to provide a practical guide for researchers. By synthesizing the current state of knowledge within this structured framework, this survey offers a clear and detailed overview for both newcomers and experts, highlighting emerging trends and future research directions in AI-driven depression detection.

The remainder of this paper is organized as follows. Section 2 defines key terms and presents foundational concepts. Section 3 introduces a structured taxonomy to categorize existing work. Section 4 provides a detailed, hierarchically structured review of the existing literature. We then introduce the notable datasets widely utilized in this domain in Section 5, while Section 6 focuses on the evaluation metrics that are critical for performance assessment. Section 8 presents the methodology used for paper selection and review, and categorization. In Section 7, we outline promising avenues for future research. Finally, Section 9 concludes the paper.

2 Preliminaries

In this section, we provide essential definitions and background information relevant to our study, grounded in established clinical and technical literature.

2.1 Depression

Depression is a common and serious medical illness that negatively affects how you feel, the way you think, and how you act [11]. It is characterized by persistent feelings of sadness and a loss of interest or pleasure in previously rewarding or enjoyable activities. Beyond these core emotional symptoms, it can also disturb sleep and appetite, and lead to tiredness and poor concentration [2]. Unlike temporary mood fluctuations in response to life’s challenges, the symptoms of clinical depression can be long-lasting and severe enough to significantly impair an individual’s ability to function at work, at school, or within the family [11, 2].

2.2 Depressive Disorder

Depressive disorder is a clinical term for a group of mood disorders where depression is the main feature. The two most common types are Major Depressive Disorder (MDD) and Persistent Depressive Disorder (dysthymia) [1]. *Major Depressive Disorder* involves discrete episodes of at least two weeks’ duration involving clear-cut changes in affect, cognition, and neurovegetative functions that cause significant distress or impairment [1]. *Persistent Depressive Disorder*, or dysthymia, is a more chronic form of depression, characterized by a depressed mood that lasts for at least two years, though it may be less severe than an episode of major depression [1]. The diagnosis and classification of these disorders are formally defined by criteria in diagnostic manuals such as the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [1].

2.3 Graph Neural Network

A Graph Neural Network (GNN) is a class of deep learning models designed specifically to perform inference on data structured as graphs [12, 13, 14, 15, 16, 17, 18]. In a graph, information is stored in nodes (entities) and edges (relationships), and GNNs leverage this structure by passing and aggregating messages between neighboring nodes [19]. This architecture allows them to effectively learn representations that capture both the features of the entities and the intricate topology of their connections. GNNs have proven highly effective in domains where relational data is key, such as social network analysis, molecular chemistry, and modeling brain functional connectivity networks [13].

2.4 Large Language Model

A Large Language Model (LLM) is a type of deep learning model, often based on the Transformer architecture, that is pre-trained on vast quantities of text data to understand, process, and generate human language [20]. Seminal models like the original Transformer introduced the self-attention mechanism, enabling models to weigh the importance of different words in a sequence [21], while subsequent models like BERT (Bidirectional Encoder Representations from Transformers) refined this by learning deep bidirectional representations from unlabeled text [22]. LLMs are capable of performing a wide array of natural language processing (NLP) tasks, including text generation, question answering, and sentiment analysis, often with minimal fine-tuning.

2.5 Electroencephalography Data

Electroencephalography (EEG) is a non-invasive neurophysiological technique used to record the electrical activity generated by the brain via electrodes placed on the scalp [23]. The resulting EEG data captures the brain’s spontaneous electrical potentials, which manifest as various rhythmic oscillations (brain waves) categorized by frequency bands such as delta, theta, alpha, and beta. These patterns reflect different states of consciousness and cognitive processes [24]. In clinical practice, EEG is a valuable tool for diagnosing and managing neurological conditions like epilepsy and sleep disorders, and it is increasingly used in psychiatric research to investigate biomarkers for conditions such as depression and schizophrenia [25].

3 A Hierarchical Taxonomy for AI in Depression Detection

To provide a clear and structured overview of the vast body of literature in AI-based depression detection, we have developed a hierarchical taxonomy. This framework organizes the 55 papers reviewed in this survey into three distinct levels: the *clinical task*, the *data type utilized*, and the *AI model employed*. This section defines the categories within each level and presents tables that classify the reviewed literature accordingly, serving as a guide for the in-depth discussion that follows in Section 4.

3.1 Classification by Task

The first level of our taxonomy distinguishes papers based on their primary clinical objective. We identify two fundamental tasks in the literature:

- *Diagnosis*: This task involves the use of AI models to identify the current depressive state of an individual. This includes classifying subjects as depressed versus non-depressed or assessing the severity of their condition based on established clinical criteria.
- *Prediction*: This task focuses on forecasting future mental health outcomes. Models are designed to predict the future onset of depression, a user’s risk level over time, or their potential response to a specific treatment.

Table 1 presents the distribution of the reviewed papers across these two tasks.

Table 1: Categorization of reviewed papers based on the primary task addressed.

Task Type	Papers
Diagnosis	[26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66]
Prediction	[67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79]

3.2 Classification by Data Type

The second level of our taxonomy categorizes studies based on the data modalities they utilize. The choice of data is a critical factor that shapes the methodological approach. We identify four main categories:

- *Text*: This modality includes written language from diverse sources such as social media platforms (Twitter, Reddit), clinical interview transcripts, and online health forums.
- *Voice*: This modality encompasses acoustic and paralinguistic features of speech, such as pitch, tone, energy, and speech rate, captured from audio recordings.
- *Neuroimaging*: This category refers to data that captures brain structure or function. Common examples in the literature include electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS).
- *Multimodal*: This approach involves the synergistic integration of two or more data types (e.g., text and audio from an interview, or EEG and eye-tracking data) to create a more holistic and robust model.

Table 2 shows the classification of the reviewed papers based on data type.

Table 2: Categorization of reviewed papers based on the data modality utilized.

Data Type	Papers
Text	[26], [28], [29], [30], [31], [32], [39], [40], [42], [70], [73], [54], [74], [55], [80], [75], [76], [61], [63], [78], [79]
Voice	[27]
Neuroimaging	[33], [35], [38], [41], [71], [72], [45], [47], [48], [49], [50], [51], [52], [57], [58], [77], [59], [66]
Multimodal	[67], [68], [34], [36], [37], [69], [43], [44], [46], [53], [56], [60], [62], [64], [65]

3.3 Classification by AI Model

The final level of our taxonomy classifies papers based on the core AI model or methodology employed. This highlights the evolution of computational techniques in the field. We define five major categories:

- *Traditional Machine Learning*: This category includes foundational models like Support Vector Machines (SVM), Naive Bayes, and Random Forests, which typically rely on manually engineered features from the data.
- *Deep Learning* : This refers to the application of a single, deep learning architecture, such as a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN) and etc.
- *Hybrid Models & Methodologies*: This class includes studies that combine multiple distinct model architectures (e.g., CNN and LSTM) to leverage their complementary strengths. It also encompasses works focused on complex methodologies like algorithmic fairness and bias mitigation.
- *Large Language Models (LLMs) & Transformers*: This category is dedicated to models based on the Transformer architecture, including seminal models like BERT and GPT, as well as their modern variants. These models are pre-trained on vast text corpora and are adept at understanding linguistic nuances.

- *Graph Neural Networks (GNNs)*: This class consists of models specifically designed to operate on graph-structured data. In this context, they are primarily used to model brain connectivity networks or social interaction graphs.

Table 3 provides a detailed breakdown of the reviewed literature according to the AI models used.

Table 3: Categorization of reviewed papers based on the primary model architecture employed.

Model Type	Papers
Traditional Machine Learning	[26], [29], [39], [70]
Hybrid Models & Methodologies	[32], [61], [52], [66], [53], [60], [62], [64], [65], [79]
Deep Learning	[28], [30], [31], [48]
LLMs & Transformers	[27], [36], [37], [69], [40], [42], [46], [73], [54], [74], [55], [80], [75], [76], [56], [63], [78]
Graph Neural Networks (GNNs)	[33], [67], [68], [34], [35], [38], [41], [43], [71], [44], [72], [45], [47], [49], [50], [51], [57], [58], [77], [59]

4 Related Work

The landscape of mental health diagnosis has evolved significantly with the advent of computational methodologies that leverage various data modalities. As traditional psychiatric practices face challenges in accurately and promptly identifying depressive disorders, researchers are increasingly turning to innovative approaches that harness the power of technology. This section reviews the related works in the field of diagnosis and prediction of depression, emphasizing the integration of textual data, voice analysis, neuroimaging, and multi-modals. By examining the diverse computational techniques employed, we highlight how these methodologies not only enhance diagnostic accuracy but also facilitate earlier intervention. The subsections that follow delve into specific modalities and methods, from traditional Machine Learning approaches to advanced deep learning architectures, illustrating the breadth of research aimed at improving mental health outcomes through systematic analysis and interpretation of data.

4.1 Depression Diagnosis

Diagnostic models focus on determining an individual’s current depressive state. These methods analyze multiple data modalities to classify subjects as depressed or non-depressed and to assess the severity of their condition according to established clinical criteria.

4.1.1 Text-Based Diagnosis

Text remains one of the most widely used data modalities, leveraging content from social media, clinical interviews, and online forums.

Traditional Machine Learning Early approaches in text-based depression detection rely on traditional machine learning models with engineered features. These foundational methods establish the viability of using linguistic patterns for mental health assessment. One prominent study [26] presents a systematic approach for detecting depression from Twitter feeds using natural language

processing (NLP). Tweets are gathered via the Twitter API with a curated list of keywords linked to poor mental well-being, requiring app authentication with consumer and access tokens. A comprehensive preprocessing pipeline—comprising link and non-ASCII removal, tokenization, stemming, stop-word elimination, and part-of-speech tagging—refines the data. Finally, the cleaned text is vectorized to train and test machine learning classifiers (support vector machines and Naive Bayes), with performance assessed by F1-score and accuracy.

In a related multi-class depression detection approach, researchers build a custom tweet dataset to predict five depression subtypes—Bipolar, Psychotic, Atypical, Postpartum, and Major Depressive Disorder. They create psychiatrist-verified lexicons of indicative phrases and use Apify to scrape matching tweets. A rigorous manual annotation process then filters the corpus to include only tweets that explicitly convey personal experiences of depression, resulting in a high-quality, contextually grounded dataset for further analysis [29].

Further advancing the generalizability of these models, another work [39] develops a depression detection model grounded in the PHQ-9 clinical questionnaire. The methodology involves developing a depression detection model grounded in the PHQ9 clinical questionnaire, which is commonly used for depression screening. The approach consists of two main components: a questionnaire model that detects symptoms from PHQ9 and a depression detection model that predicts depression based on symptom presence in social media posts. The models range from rule-based pattern matching to a BERT-based classifier, progressively relaxing constraints to allow more flexibility in learning. The dataset consists of three Reddit-based depression detection datasets: RSDD, eRisk2018, and TRT. These datasets differ in construction methodologies, such as self-reported diagnoses or participation in mental health forums. A weakly supervised approach is used to label symptoms by leveraging regular expressions, sentiment models, and heuristics, ensuring the model generalizes well across datasets while remaining interpretable.

Hybrid Models & Methodologies Several studies combine different deep learning architectures to leverage their complementary strengths for text-based diagnosis. One such work [32] presents a method for detecting depression through the analysis of social media text using a hybrid deep learning model known as Fasttext Convolution Neural Network with Long Short-Term Memory (FCL). This approach aims to improve early detection of depression, which is critical for timely intervention. The FCL model leverages Fasttext embeddings to enhance word representation, addressing limitations of traditional methods by capturing semantic information and out-of-vocabulary words. It combines Convolutional Neural Networks (CNN) for global feature extraction and Long Short-Term Memory (LSTM) networks for understanding local dependencies in text. The methodology involves data cleaning and preprocessing of a dataset containing tweets/posts labeled for depression, followed by the application of the FCL model, which includes padding strategies to ensure consistent input sizes. The model’s performance is evaluated against existing state-of-the-art methods, demonstrating higher accuracy in detecting depression from social media content.

In [61], the Depressive Emotion–Context Enhanced Network (DECEN) detects depression from about 40 000 Sina Weibo posts—including a DSM-5–annotated subset of term-level depressive emotions. DECEN comprises: (i) a BiLSTM+CRF Depressive Emotion Recognition module for semantic–syntactic extraction; (ii) a BERT+attention Emotion-Context Enhanced Representation module; and (iii) a BiLSTM classification layer. By modeling specific depressive emotions (e.g., anhedonia, suicidal ideation) and their contextual relations, DECEN outperforms generic sentiment or embedding-based models in accuracy, precision, and robustness to sarcasm and variable text lengths.

Deep Learning With the advent of deep learning, researchers begin to utilize architectures like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to automatically learn feature representations from text. A notable study [28] investigates the identification and intensity assessment of depression through tweets on Twitter, employing artificial intelligence and deep learning techniques, specifically long short-term memory (LSTM) models. The dataset comprises 95,322 tweets labeled as “non-depressed” or “depressed,” with the latter further classified into “mild,” “moderate,” and “severe” categories based on clinical criteria from the DSM-5. Data collection focuses on tweets containing specific hashtags associated with depressive symptoms while excluding uplifting or irrelevant content. Emotional and semantic scores are calculated using the Valence Aware Dictionary and Sentiment Reasoner (VADER) and latent semantic indexing (LSI) methods, respectively. The final depression intensity score is normalized and categorized into four classes.

In [30], Patil et al. develop a system to detect and assess depression levels from medical data by analyzing unstructured text with a convolutional neural network (CNN). To address the lack of a large, publicly available benchmark for depression analysis, the pipeline begins with text preprocessing and proceeds to feature extraction and classification. The model treats text as a one-dimensional matrix and employs one-dimensional convolutional layers to capture local patterns. A word embedding layer first transforms tokens into dense vectors, which the convolutional network then processes to extract salient features. Pooling layers reduce dimensionality while retaining key information, and fully connected layers integrate these features into a unified representation. Finally, a SoftMax classifier predicts depression levels based on the learned features [30].

In [31], Liu et al. propose DeCapsNet, which combines capsule networks with contrastive learning for interpretable, high-performance depression detection from online posts. The model overcomes prior limitations by extracting symptom capsules based on PHQ-9 descriptions, enabling hierarchical reasoning. DeCapsNet first assigns each post a risk score—computed by its similarity to PHQ-9 symptoms—to select the most relevant samples. Capsule layers then generate symptom and depression capsules, aggregating symptom features into a higher-level representation for classification. A contrastive learning objective further refines embeddings by pulling similar instances together and pushing different classes apart. Evaluated on eRisk2018, RSDD, and TRT datasets, DeCapsNet outperforms baseline models in both within- and cross-dataset settings. Its loss function combines dynamic routing and contrastive learning losses to enhance overall performance [31].

LLMs & Transformers The current state-of-the-art in depression detection is dominated by Large Language Models (LLMs) and Transformer architectures, which leverage pre-trained knowledge to capture the nuanced, contextual nature of language. Chen et al. [36] propose a Structural Element Graph (SEGA) for depression detection in clinical interviews. SEGA employs a directed acyclic graph to model interactions among interview components—questions, transcripts, audio, and video—structured according to expert knowledge to ensure meaningful feature extraction while reducing noise. To address data scarcity, the authors introduce an LLM-guided data augmentation strategy that generates synthetic responses and rephrased transcripts. Contrastive learning is then applied to refine representations by drawing similar instances closer and separating dissimilar ones. They evaluate SEGA on two real-world multimodal clinical interview corpora: DAIC-WOZ (English interviews labeled with PHQ-8) and EATD (Chinese interviews labeled with SDS). Although both datasets are small due to privacy constraints, augmenting them with LLM-generated responses enhances training robustness and improves depression detection performance [36].

Similarly, Zhang et al. [37] integrate speech signals into Large Language Models using Acoustic

Landmarks. Their method proceeds in three stages: (1) extracting phonetic landmarks from raw speech, (2) fine-tuning the LLM with cross-modal instructions to interpret these landmarks, and (3) applying P-Tuning for depression classification. They evaluate on the DAIC-WOZ corpus of clinical interviews and mitigate data scarcity and class imbalance through sub-dialogue shuffling. By combining text inputs with acoustic landmarks, the approach achieves a state-of-the-art F1 score of 0.84, demonstrating that lightweight landmark features can rival more complex deep-learning-based speech representations. Other work applies transformer networks—such as BERT, GPT-3.5, and ChatGPT-4—to clinical interviews by augmenting traditional datasets with simulated data to preserve privacy and boost performance [40]. Using DAIC-WOZ and Extended-DAIC, these studies focus on text-based modalities, preprocess inputs to address class imbalance

In [42], the authors propose a language-model-only framework for automated depression assessment using the Extended DAIC-WOZ (E-DAIC) corpus, which features semi-clinical interviews. The pipeline begins with data preprocessing, where OpenAI’s Whisper model transcribes incomplete or noisy audio into accurate text. Feature extraction consists of two stages: first, GPT-3.5-Turbo transforms the interview transcripts to highlight depression-relevant patterns; second, DepRoBERTa—a RoBERTa model fine-tuned on PHQ-labeled data—classifies depression severity. Transcripts annotated with PHQ scores are then used to train a Support Vector Regression (SVR) model, with hyperparameters optimized via GridSearchCV. This methodology demonstrates how advanced NLP techniques can improve the accuracy of automated depression detection and establishes a foundation for future research in the field [42].

Other research [40] applies transformer-based networks—including BERT, GPT-3.5, and ChatGPT-4—to clinical interview transcripts. To address privacy concerns and augment limited corpora, the authors generate simulated data, improving depression recognition from linguistic features. While the framework can handle multimodal inputs, ethical constraints restrict evaluation to text, leveraging PHQ-8 scores for ground-truth labeling. The study details a comprehensive preprocessing pipeline—covering transcript cleaning, tokenization, and class-balancing—and achieves state-of-the-art performance in automated depression detection. These findings underscore AI’s transformative potential for early mental health intervention and pave the way for future diagnostics [40].

Zhang et al. [55] propose RED (Retrieval-augmented generation for Explainable Depression detection), a transparent method for depression detection from clinical interview transcripts. RED relies solely on text modality and grounds its predictions in retrieved evidence to prevent hallucinations. The model is evaluated on DAIC-WoZ, a structured corpus of interview transcripts labeled with PHQ-8 scores. It employs an adaptive Retrieval-Augmented Generation framework, first generating personalized queries via an LLM to infer each participant’s profile and then retrieving context-sensitive dialogue segments. Retrieved evidence is enriched by a social intelligence module that integrates relevant knowledge from the COKE cognitive knowledge base using an event-centric retriever. This augmented evidence supports both the depression prediction and the generation of interpretable explanations. By tailoring retrieval to individual contexts and incorporating psychologically relevant knowledge, RED outperforms neural and post-hoc LLM baselines in accuracy and explanation quality. These results highlight RED’s potential for trustworthy, personalized mental health assessments [55].

In [42], the authors introduce a language-model-only framework for automated depression detection using the Extended DAIC-WOZ (E-DAIC) corpus. E-DAIC comprises semi-clinical interviews conducted by the virtual agent Ellie in a Wizard-of-Oz setting, involving 275 participants (170 males, 105 females) split into 163 training, 56 development, and 56 test instances. Each interview, lasting approximately 20 minutes, is labeled with a PHQ-8 score to assess depression severity. The

pipeline begins with data preprocessing: Whisper transcriptions of incomplete or noisy audio are corrected using OpenAI’s Whisper model. Feature extraction employs a two-stage strategy to prevent overfitting. First, GPT-3.5-Turbo transforms transcripts by generating prompts that highlight depression-relevant features. Second, DepRoBERTa—a RoBERTa variant fine-tuned on these enhanced transcripts—classifies depression severity into “none”, “moderate”, or “severe”. Finally, the PHQ-8 annotations guide the training of a Support Vector Regression (SVR) model to predict continuous PHQ scores, with hyperparameters optimized via GridSearchCV.

In [54], Teng et al. introduce a novel Chain-of-Thought (CoT) prompting strategy for Large Language Models (LLMs) to improve interpretability and clinical alignment in text-based depression detection. The approach uses the E-DAIC dataset, which comprises PHQ-8-annotated transcribed clinical interviews. The detection task is decomposed into four structured reasoning stages:

1. Emotion analysis to identify affective states.
2. Binary classification to detect the presence of depression.
3. Causal reasoning to uncover contributing and protective factors.
4. Severity assessment aligned with PHQ-8 scoring.

This explicit reasoning framework mirrors clinical diagnostic workflows and captures nuanced symptoms, such as anhedonia expressed via negated positives. CoT prompts are applied to multiple LLMs—including GPT-4o, Qwen2.5, and DeepSeek R1—and yield significant improvements in concordance correlation coefficient (CCC) and mean absolute error (MAE). Notably, GPT-4o with CoT achieves $CCC = 0.732$ and $MAE = 3.37$. Ablation studies confirm that explicit CoT prompting enhances even inherently reasoning-capable models, underscoring its value for reliable, interpretable mental health assessments [54].

4.1.2 Voice-Based Diagnosis

Voice data contain rich paralinguistic and acoustic cues that strongly indicate depressive states.

LLMs & Transformers Voice-based analysis has greatly benefited from pre-trained models capable of extracting high-quality features from raw audio. A recent paper introduces a novel artificial-intelligence approach for the early detection of depression using voice data, addressing the challenge of limited dataset size. The methodology employs the wav2vec 2.0 pre-training model as a feature extractor to derive high-quality voice representations from raw audio, specifically utilizing the DAIC-WOZ dataset, which comprises 189 dialogue recordings between patients and a virtual agent. The recordings are preprocessed to enhance data quality through voice segmentation, ensuring that only patient speech segments are analyzed. The audio files are randomly divided into training, validation, and test sets in a 6:2:2 ratio. The wav2vec 2.0 model, renowned for its effectiveness in speech processing, is fine-tuned to classify depression, achieving accuracies of 0.9649 for binary classification and 0.9481 for multi-class classification. The model architecture includes a feature encoder for capturing local acoustic patterns and a transformer module for modeling global context, with additional dropout layers to mitigate overfitting during training. Overall, the proposed method demonstrates strong generalization capabilities and practical applicability for assisting healthcare professionals in early depression screening [27].

4.1.3 Neuroimaging-Based Diagnosis

Neuroimaging modalities, such as EEG and fMRI, offer direct insights into the neural correlates of depression, while graph neural networks (GNNs) excel at analyzing brain connectivity patterns.

GNNs GNNs have emerged as the dominant architecture for modeling the non-Euclidean structure of brain networks. One pioneering study introduced a more accurate and objective method for detecting depression, moving beyond traditional clinician–patient assessments. This approach uses resting-state electroencephalography (EEG) recordings from 27 patients with depression and 28 healthy controls to construct a brain functional network. Functional connectivity matrices are generated using Pearson correlation, from which four linear EEG features—activity, mobility, complexity, and power spectral density—are extracted. The authors propose the Graph Input Layer Attention Convolutional Network (GICN), which integrates a trainable weight matrix in the input layer and adopts the brain functional network as its adjacency matrix. In 10-fold cross-validation, GICN achieves a recognition accuracy of 96.50%, outperforming existing methods. The model also identifies critical brain regions—specifically the temporal and parietal–occipital lobes—that significantly contribute to depression classification. Their methodology includes rigorous EEG acquisition and preprocessing to ensure high-quality data, and the use of graph convolutional layers to enhance feature extraction and classification [33].

Another study presents a novel depression recognition method using a graph neural network that integrates spatio-temporal features extracted from functional near-infrared spectroscopy (fNIRS) data. The authors collected fNIRS recordings from 96 participants and derived six statistical metrics—mean, maximum, minimum, variance, skewness, and kurtosis—as temporal node features, alongside channel connectivity measures (coherence and correlation) as spatial edge weights. Each subject’s data is modeled as a graph where nodes encode temporal attributes and edges represent spatial relationships, enabling the GNN to learn jointly from both feature types. Experimental results demonstrate that this approach surpasses traditional machine learning methods in accuracy, F1 score, and precision, achieving over a 10% improvement in F1 score. The dataset, acquired at Renmin Hospital of Wuhan University, comprises 53-channel fNIRS signals sampled at 100 Hz during a verbal fluency task. This study underscores the efficacy of combining temporal and spatial information for automatic depression recognition [35].

In [38], the BrainIB framework is introduced—a graph neural network that leverages the information bottleneck (IB) principle to improve psychiatric disorder diagnosis from fMRI-derived connectivity data. Traditional classifiers often overfit and lack explainability, limiting their clinical utility. BrainIB overcomes these challenges by sampling informative subgraphs from whole-brain functional connectivity graphs, thereby enhancing generalization to unseen data. Evaluated on three psychiatric cohorts—ABIDE, REST-meta-MDD, and SRPBS—BrainIB outperforms seven state-of-the-art methods in diagnostic accuracy. fMRI data are preprocessed with slice-timing correction, motion correction, and normalization, followed by extraction of functional connectivity (FC) matrices. The BrainIB architecture comprises:

1. A subgraph generator that samples informative subgraphs from FC graphs.
2. A graph encoder that learns embeddings for each subgraph.
3. A mutual information estimator that quantifies the dependence between subgraphs and the original graphs.

In addition to its superior accuracy, BrainIB identifies subgraph biomarkers consistent with clinical findings, underscoring its potential for practical psychiatric applications [38].

Other research has focused on fusing static and dynamic brain networks. The Dynamic-Static Fusion Graph Neural Network (DSFGNN) is designed to diagnose Major Depressive Disorder (MDD) by integrating static and dynamic functional connectivity networks derived from resting-state fMRI data. It addresses challenges in brain connectivity modeling, representation learning, and interpretability. Specifically, static functional connectivity graphs are constructed using Pearson correlation coefficients computed over the entire fMRI time series, while dynamic graphs are generated via a sliding-window approach to capture temporal fluctuations. The methodology employs separate Graph Isomorphism Network encoders for the static and dynamic graphs to learn node-level representations, a spatiotemporal attention mechanism to aggregate these representations into a global graph embedding, and a temporal attention module with Gated Recurrent Units (GRU) to model temporal evolution. Additionally, DSFGNN incorporates causal disentanglement to isolate causal factors from non-causal ones—enhancing interpretability—and applies orthogonal regularization to promote diverse representations for improved generalization and stability. The framework is evaluated on the Rest-meta-MDD dataset, a multi-site resting-state fMRI collection originally comprising 1,300 MDD patients and 1,128 healthy controls across 25 sites. After quality control and site selection, 832 MDD patients and 779 controls from 17 sites remain. All participants provided informed consent, and the study received ethical approval. Preprocessing steps included slice-timing correction, head-motion correction, nuisance regression, spatial normalization, and band-pass filtering (0.01–0.1 Hz) [41].

Further research has explored multi-view learning and advanced graph neural network architectures. The Multi-View Graph Neural Network (MV-GNN) integrates spatial and topological information from functional connectivity (FC) networks derived from resting-state fMRI. This study leverages the REST-meta-MDD dataset, which includes data from 1,300 Major Depressive Disorder (MDD) patients and 1,128 healthy controls (HC) collected across 25 research groups in 17 hospitals throughout China [45]. The methodology extracts complementary views from FC networks to elucidate MDD’s neural correlates. Spatial connectivity patterns altered in MDD are identified via T-tests and refined using LASSO for feature selection while preventing information leakage. Model interpretability is enhanced through SHAP analysis, which underscores the cerebellum’s significance. Additionally, the topology of FC networks is examined at both global and local scales, revealing structural differences between MDD and

In a similar vein, the DepressionGraph framework employs a two-channel graph neural network (GNN) alongside a transformer-based architecture to capture time-varying information from brain functional connectivity networks. The model leverages the REST-meta-MDD consortium’s publicly available resting-state fMRI dataset, which comprises 533 subjects across 17 Chinese hospitals [47]. For each subject, the fMRI time series is divided into discrete time slices, and functional connectivity networks (FCNs) are constructed where nodes correspond to brain regions of interest (ROIs) and edges reflect correlation coefficients between node features. A gated recurrent unit (GRU) network first encodes node features to integrate temporal context. These encoded features are then processed by a two-channel GNN: one channel extracts fine-grained local connectivity patterns, while the other captures coarse-grained global structures. A transformer module subsequently models temporal dynamics across the sequence of FCNs. Final classification into MDD or control is achieved with

Other studies have introduced graph autoencoders (GAEs) for brain disorder diagnosis. This framework leverages graph convolutional networks to perform inductive embedding of functional connectivity (FC) networks derived from resting-state fMRI data. By preserving the non-Euclidean

structure of brain graphs, it departs from traditional convolutional neural network approaches. First, individualized FC graphs are constructed using the Ledoit–Wolf shrinkage estimator, followed by extraction of node-level features. A multi-layer GCN encoder then aggregates local neighborhood information via spectral convolution. In the unsupervised variant, a symmetric decoder reconstructs the original adjacency matrix from learned embeddings. In the supervised variant, the GCN encoder is trained end-to-end with a fully connected neural network (FCNN) that maps graph-level embeddings directly to diagnostic labels. Because the GCN encoder is inductive, it generalizes across subjects’ networks without relying on a fixed population graph [49].

An ensemble graph neural network model has also been proposed for MDD diagnosis. This study leverages resting-state fMRI data from the REST-meta-MDD collaboration, comprising 1,586 participants. Functional connectivity matrices are generated for each subject and represented as graphs, where nodes correspond to brain regions of interest (ROIs) and edges encode pairwise connectivity strengths. To boost classification performance, the framework ensembles three base GNN architectures—Graph Convolutional Network, Graph Attention Network, and GraphSAGE—each processing the same graph input to learn complementary embeddings. The learned representations are concatenated and fed into a meta-classifier with softmax activation to produce the final prediction. By combining diverse modeling biases, this ensemble approach improves generalization and accuracy over individual models, enabling discrimination between MDD and healthy controls as well as between first-episode drug-naïve (FEDN) and recurrent (REC) MDD subtypes [50].

The Adaptive Propagation Operator Graph Convolutional Network (APO-GCN) introduces an adaptive propagation operator to mitigate over-smoothing and better capture discriminative patterns in brain functional graphs. This framework is applied to resting-state fMRI data from the multi-site REST-meta-MDD Consortium. After constructing individual functional connectivity matrices from preprocessed BOLD signals, APO-GCN employs Chebyshev polynomial-based convolutions for computational efficiency and dynamically adjusts its propagation operator during training. This adaptive modulation of information flow between graph nodes preserves critical signals that conventional GCNs tend to homogenize. Model performance is validated using both 10-fold cross-validation and leave-one-site-out schemes, achieving classification accuracies up to 91.8%. Interpretability is further enhanced through GNNExplainer, which identifies the most influential brain regions driving MDD classification [51].

Further advancements have introduced DSGNN, a dual-branch self-supervised graph neural network (GNN) that leverages contrastive learning for depression diagnosis using resting-state fMRI data. In this framework, brain functional connectivity networks are modeled as graphs, and a self-supervised architecture comprising two parallel GNN branches is proposed. These branches are trained to maximize agreement between differently augmented views of the same graph via a contrastive loss. Data augmentations—perturbing node features or graph structure—encourage the model to learn invariant and robust representations. Each branch consists of a GCN-based encoder followed by a projection head, and the resulting embeddings are contrasted in the latent space. After self-supervised pretraining, the shared encoder is fine-tuned for the downstream diagnostic classification task using a multilayer perceptron [57].

Further research has introduced a Frequency Feature Fusion Graph Network leveraging functional near-infrared spectroscopy (fNIRS) data for depression diagnosis. A new dataset comprising 1,086 subjects was collected to support this study. The proposed model is built on a Temporal Graph Convolutional Network (TGCN) that integrates spatial and temporal brain features extracted from fNIRS signals. Temporal dynamics are enriched via a Discrete Fourier Transform (DFT), enabling the identification of frequency-domain biomarkers. These frequency features are

then fused with original spatial features through a three-stage, phase-specific GCN architecture that independently models the silent (resting), task, and post-task periods. Key innovations include a Temporal Fusion Module (TFM) that merges raw and frequency-based representations, and a Frequency Point-Biserial Correlation Attention Module (FAM) that assigns attention weights to the most discriminative channels and frequency bands for improved diagnostic accuracy [58].

Finally, one study applies graph neural networks to causal connectomes derived from resting-state fMRI (rs-fMRI) in 1,296 young adults [59]. The methodology develops and compares GNN-based classifiers using both traditional functional connectomes—built from Pearson and partial correlations—and three causal connectome methods: the TwoStep algorithm, Granger causality, and regression dynamic causal modeling (rDCM). These graphs are processed by advanced GNN architectures, including MSGNN, GIN, and GAT. The principal innovation is that causal connectivity not only outperforms functional connectivity in classification but also enhances neurobiological interpretability. Through GNNExplainer and spatial correlations with PET ligand maps, the study links model-identified key nodes to neurotransmitter systems such as serotonin (5-HT1B) and dopamine (extrastriatal D2) [59]. Another contribution is the node-aware contrastive graph learning (NCGL) framework, which advances self-supervised GNNs by modeling individual functional connectivity networks as graphs [63]. NCGL’s core innovation is a contrastive pretraining strategy that incorporates node-level discrimination into graph-level contrastive learning. It employs dual GCN encoders to process two augmented views of the same graph and aligns their embeddings with a hybrid loss: a graph-level contrastive term plus a node-level consistency regularizer. The node-level loss penalizes discrepancies in corresponding node embeddings across augmentations, enforcing fine-grained invariance.

Hybrid Models & Methodologies Some studies have combined diverse modeling approaches to capture complementary aspects of neuroimaging data. For example, the Hybrid Graph Neural Network (HybGNN) framework leverages EEG signals for depression detection using a dual-branch architecture comprising a Common Graph Neural Network (CGNN) and an Individualized Graph Neural Network (IGNN). The CGNN employs a fixed graph topology to learn shared depression-related patterns, whereas the IGNN dynamically constructs subject-specific graphs to capture individual abnormalities. To enhance hierarchical feature learning, HybGNN integrates a Graph Pooling and Unpooling Module (GPUM) that adaptively aggregates EEG channels into brain regions and reinjects this structural information back into the network. Evaluated on the publicly available MODMA and HUSM EEG datasets, HybGNN outperforms state-of-the-art models, achieving accuracies of 95.42% on MODMA and 93.50% on HUSM [52].

Another significant contribution is the systematic investigation of machine learning fairness in EEG-based depression detection. This study represents the first comprehensive analysis of algorithmic bias in depression classification using electroencephalography (EEG) data. Three benchmark EEG datasets—Mumtaz, MODMA, and Rest—are evaluated; they differ in gender distribution, sampling rates, and electrode configurations. The methodology involves training three deep learning models for depression classification: Deep-Asymmetry (CNN-based), GTSAN (GRU with attention), and 1DCNN-LSTM (convolutional-recurrent network). Five bias mitigation strategies are applied at various stages: pre-processing (Mixup augmentation, data massaging), in-processing (loss reweighting, regularization for equalized odds), and post-processing (Reject Option Classification). The innovation lies in the comprehensive fairness-aware framework, which employs metrics such as statistical parity, equal opportunity, equalized odds, and equal accuracy to quantify bias and examine its sensitivity to model architecture, dataset characteristics, and mitigation techniques.

The findings reveal persistent algorithmic and dataset biases across all models and datasets, with pronounced disparities along gender lines [66].

Deep Learning The SGP-SL model [48] introduces a three-stage framework for detecting Major Depressive Disorder (MDD) from EEG signals: graph construction, self-attention graph pooling, and prediction. First, an adjacency matrix is constructed to represent relationships between EEG electrodes, capturing both local and global connectivity patterns. Next, this graph is refined through multiple self-attention graph pooling modules, which preserve critical information while reducing the graph to a unified vector representation. Finally, the pooled representation is fed into a multi-layer perceptron (MLP) that simultaneously predicts soft class labels and Patient Health Questionnaire-9 (PHQ-9) scores. The loss function combines classification and regression objectives, using Kullback–Leibler divergence for the classification loss, mean absolute error for the regression loss, and an additional disagreement loss term to strengthen the correlation between tasks. The model is evaluated on the MODMA dataset, which comprises resting-state EEG recordings from 53 subjects (24 MDD patients and 29 healthy controls) acquired with 128 electrodes [48].

4.1.4 Multimodal Diagnosis

Multimodal approaches combine data from diverse modalities—such as text, audio, visual, and physiological signals—to develop more comprehensive and robust diagnostic models.

GNNs Graph neural networks excel at fusing features from multiple modalities represented within graph structures. The Local-Global Multimodal Fusion Graph Neural Network (LGMF-GNN) [34] integrates functional MRI, structural MRI, and electronic health records (EHRs) to improve the objectivity of Major Depressive Disorder (MDD) diagnosis. Tested on diverse, multinational cohorts, the LGMF-GNN achieved a classification accuracy of 78.75% and an AUROC of 80.64%, effectively distinguishing MDD subtypes and uncovering distinct brain connectivity patterns associated with the disorder. In the proposed local-global architecture, we begin by constructing region-of-interest (ROI) graphs for each subject. A learnable adjacency matrix is derived from the ROI BOLD time series, and node attributes correspond to the columns of the functional connectivity matrix obtained from resting-state fMRI. A local ROI GNN then applies graph convolution with an attention mechanism to aggregate ROI information, while a gated recurrent unit (GRU) encoder generates regional embeddings from the time series. These embeddings feed into a graph generator that outputs a subject-specific adjacency matrix. The GNN predictor uses attention over this learned graph and its node features to produce local embeddings and classification results.

To incorporate anatomical and demographic data, we designed a Global-Local Transformer (GLT) encoder and a Pairwise Association Encoder (PAE), which transform T1-weighted MRI features and demographic variables into one-dimensional feature vectors. In the global subject GNN, functional, anatomical, and demographic features serve as node attributes across three subject graphs. At the population level, modality-specific GCN blocks generate representations unique to each modality, while modality-common GCN blocks distill shared information. A multimodal attention block then refines these representations into a unified embedding. Finally, a multilayer perceptron (MLP) classifier produces the global prediction [34].

The MS²-GNN model [43] enhances Major Depressive Disorder (MDD) detection by integrating EEG and audio signals from the MODMA dataset. It comprises several key components designed for effective multimodal feature extraction and fusion. Initially, Long Short-Term Memory (LSTM)

networks extract task-oriented features from both audio and EEG modalities, capturing MDD-related characteristics while reducing dimensionality. A shared network then identifies common representations across modalities, and modality-specific networks encode unique features for each. The embeddings are reconstructed to preserve semantic integrity, and an attention mechanism fuses them into a compact multimodal representation for classification. The dataset consists of multimodal samples with audio and EEG inputs denoted X_a and X_e , respectively, alongside ground-truth labels. A GNN refines these representations by propagating information over a dynamically constructed affinity matrix, computed from the absolute differences between node embeddings. This strategy enables the model to learn inter-modality relationships without requiring a predefined adjacency structure [43].

The KARE framework integrates physical activity data from smart home sensors and cyberspace activity (e.g., internet logs) to detect early signs of depression in older adults. This study proposes the KARE (Knowledge graph-based Activity pattern Recognition for Early detection of depression) framework. The task is formulated as an anomaly detection problem, identifying deviations in individuals’ activity patterns that may signal depression. The methodology involves constructing a knowledge graph (KG) to integrate heterogeneous data sources, resolving semantic inconsistencies and enabling a unified representation. The Cyber-Physical View Representation (CPVR) module maps both physical and cyberspace activities into the KG, while the Personalized Activity Pattern Recognition (PAPR) module employs a Graph Attention Network (GAT) to learn normal behavioral patterns and detect anomalies. The innovation lies in the cross-domain fusion of data sources through graph-based representation learning, enhancing the accuracy and timeliness of depression detection in smart home environments [44].

LLMs & Transformers LLMs can serve as a powerful backbone for fusing multimodal data. The DSE-HGAT model [46] detects depression from clinical interview transcripts by constructing a heterogeneous graph. The task is framed as a dialogue extraction problem, where each transcript comprises a sequence of questions and responses labeled as “depressed” or “non-depressed.” The model consists of three primary components:

1. Context encoder layer: Two BiLSTM networks capture both local utterance-level and global dialogue-level information. Semantic features (word embeddings) and syntactic features (part-of-speech tags and named entity recognition embeddings) are integrated.
2. Heterogeneous graph layer: A graph is constructed with five node types—word, utterance, speaker, type, and state—and four edge types representing relationships within the interview. A graph attention mechanism aggregates information across these nodes to model contextual dependencies.
3. Output layer: Representations of word and state nodes are aggregated and fed into a classifier to predict depression labels. To address class imbalance, a focal loss function is employed during training.

The model is evaluated on the DAIC-WOZ dataset [46].

SpeechT-RAG [56] is a novel framework for depression detection that integrates acoustic and textual modalities, with a focus on speech-derived temporal features. This approach is evaluated on the DAIC-WOZ corpus and addresses shortcomings of text-only large language models (LLMs) and retrieval-augmented generation (RAG) systems, which often fail to capture critical nonverbal cues. The methodology begins by extracting acoustic landmarks—such as glottal onsets and voiced

frications—from the speech signal. Durations between successive landmarks are computed to form temporal bigrams, which are then summarized into statistical features (e.g., mean, variance). These timing-based features serve as retrieval keys in a novel RAG process that selects relevant examples without additional fine-tuning. Retrieved examples are converted into structured prompts and fed to LLMs (LLaMA2, LLaMA3) for classification. A Gaussian Process Classifier provides calibrated confidence estimates. By leveraging speech timing patterns as retrieval keys, SpeechT-RAG achieves higher F1 scores and lower calibration errors compared to text-only RAG and traditional fine-tuned models [56].

Hybrid Models & Methodologies Hybrid models are particularly powerful in multimodal settings, as they can combine specialized architectures for each data type. A two-stage graph neural network methodology is proposed for depression detection using audio signals [53]. The model is evaluated on three diverse datasets—DAIC-WOZ, MODMA, and D-Vlog—that include clinical interviews and real-world recordings. First, low-level frame-wise audio features (MFCCs, log F0, and constant-Q transform coefficients) are extracted and passed through a gated recurrent unit (GRU) network to capture temporal dependencies. Next, a two-stage GNN is employed:

1. Intra-audio graph: each audio frame is treated as a node, and graph attention layers aggregate contextual frame-level information.
2. Inter-audio graph: each node represents a full audio embedding, with edges weighted by cosine similarity and refined via an emotion-aware attention mechanism.

A pre-trained CompactSER model generates high-level sentiment features, which are integrated with the GNN embeddings using a self-attention fusion module. This hybrid, hierarchical graph-based architecture significantly outperforms traditional methods [53].

Similarly, the Multimodal Object-Oriented Graph Attention Model (MOGAM) is a deep learning framework for detecting depression in social media vlogs by integrating visual, textual, and structural modalities. The dataset, curated from YouTube, comprises 4,767 vlogs categorized into daily, high-risk depression, and clinically diagnosed depression groups. For each vlog, YOLOv5 identifies objects in video frames to construct an object co-occurrence graph, whose adjacency matrix is processed by a graph neural network (Graph Convolutional Network, Graph Attention Network, or GraphSAGE). The resulting graph features are fused with visual embeddings from a ResNet backbone and textual embeddings from KoBERT. A cross-attention mechanism within a transformer architecture integrates these modalities into a unified representation. MOGAM’s object-oriented approach avoids reliance on human-centric cues such as facial expressions or body poses, thereby improving generalizability [60].

The Multimodal Transformer Network (MTNet) [62] is designed for mild depression detection by integrating electroencephalography (EEG) and eye-tracking data. The dataset comprises recordings from 49 adolescents, including 21 with mild depression and 28 healthy controls. EEG signals are preprocessed into frequency-specific segments, then passed through spatial and temporal convolutional layers before being encoded by a multi-head self-attention transformer. Eye-tracking features are extracted from fixation patterns and incorporated at various fusion stages—early, intermediate, and late—between the two modalities. The key innovation of MTNet lies in its exploration of fusion timing and method, demonstrating that intermediate fusion achieves the highest classification accuracy of 91.79% [62].

The lightweight cross-modality model [64] combines audio and text data to detect depression across three multilingual datasets: DAIC-WOZ (English), EATD-Corpus (Chinese), and the Korean Depression Dataset. Audio signals are converted into Mel spectrograms and processed by an MLP-Mixer, while textual transcripts are encoded using an XLM-RoBERTa transformer encoder. The resulting embeddings are fused via a cross-attention mechanism that enables dynamic, bidirectional interaction between the two modalities. A key innovation is the model’s streamlined architecture, which significantly reduces parameter count without sacrificing accuracy. Furthermore, the approach demonstrates strong cross-lingual generalizability across all three languages [64].

The U-Fair framework [65] is an uncertainty-based multimodal multitask learning model that leverages audio, visual, and textual modalities from the DAIC-WOZ and E-DAIC datasets for fairer depression detection. Grounded in the clinical structure of the PHQ-8 questionnaire, U-Fair treats each of its eight symptom-specific items as a separate task within a multitask learning setup. To mitigate demographic bias, U-Fair introduces a gender-aware loss reweighting strategy based on aleatoric uncertainty. This dynamic reweighting adjusts task priorities according to observed task difficulty and gender-specific symptom distributions. The framework’s principled integration of gender-based uncertainty into loss optimization aligns with clinical diagnostic logic and improves fairness in model predictions [65].

4.2 Depression Prediction

Prediction models are designed to forecast the onset of depression or estimate a user’s risk level over time. These models are crucial for enabling early intervention and preventative care.

4.2.1 Text-Based Prediction

Analyzing longitudinal text data can reveal patterns predictive of future depressive episodes.

Traditional Machine Learning A depression prediction model for Arabic social media posts is developed using a psychologist-annotated Twitter dataset of 1,058 tweets, evenly divided into “depressed” and “non-depressed” classes [70]. The proposed framework comprises four phases: i) Preprocessing: cleaning and normalizing raw tweets. ii) Medical concept extraction: mapping texts to UMLS concepts via quickUMLS. iii) Feature representation: weighting extracted concepts with Bag-of-Words (BOW) and TF-IDF to generate numerical vectors. iv) Classification: training five machine learning algorithms—Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD).

Hybrid Models & Methodologies DEPRESSIONX is an explainable deep learning model for detecting depression severity from social media text. We evaluate it on two benchmark Reddit datasets (D1 and D2), where posts are labeled into four ordered severity levels—minimal, mild, moderate, and severe—according to DSM criteria and the Beck Depression Inventory (BDI) standards [79]. The DEPRESSIONX architecture comprises:

1. Multi-level textual encodings (word-, sentence-, and post-level).
2. Knowledge infusion via a structured depression knowledge graph built from Wikipedia using the REBEL relation extraction model.

3. Residual multi-head attention to fuse textual vectors with knowledge graph embeddings.
4. GIN-GAT layers for modeling structural relationships within the knowledge graph.

Severity predictions are produced through ordinal regression to respect the inherent ordering of depression classes. Explainability is achieved by: i) Attention maps that highlight critical words and sentences, and ii) Visualization of optimized knowledge subgraphs revealing the key conceptual relationships driving each prediction.

LLMs & Transformers Large Language Models (LLMs) are increasingly applied to predictive tasks on social media. Qin et al. [73] introduce an interactive, explainable system for depression detection using the TMDD (Twitter) and WU3D (Weibo) datasets. Their methodology represents each user as a collection of posts containing both text and images, and trains a vanilla depression detection model, F^* , on the full dataset to generate predicted depression probabilities that serve as answer heuristics. To enhance interpretability, the system employs a Chain-of-Thought (CoT) framework that articulates the reasoning behind each diagnosis in a structured question-answer format. Additionally, a tweet selector filters relevant posts, an image descriptor converts visual content into textual descriptions, and a prompt manager interfaces with LLMs to deliver diagnostic results and facilitate interactive dialogue [73].

Another study fine-tunes GPT-3.5 Turbo 1106 and LLaMA2-7B on a large-scale dataset of 40,000 Twitter posts, which comprises three subsets: a depression-labeled set (D1), a non-depression set (D2), and a depression-candidate set (D3). The methodology involves carefully fine-tuning the pre-trained LLMs with optimized hyperparameters to distinguish depressive from non-depressive content by leveraging linguistic cues and emoji sentiment analysis. This domain-specific refinement enables real-time, scalable monitoring of mental health signals on social media. The fine-tuned GPT-3.5 Turbo achieves a detection accuracy of 96.4%, while LLaMA2-7B reaches 87.1% [74].

In a different direction, Kumar et al. [78] develop the AST-D system (Abstractive Summarization Transformer for Depression) to automate the summarization of depression detection literature. This work addresses the ever-growing volume of research in depression detection by proposing an AI-powered pipeline that converts full-text scientific articles into concise abstracts. The methodology leverages DepressiLex, a dataset of 40 peer-reviewed papers, and fine-tunes multiple pre-trained transformer architectures—T5-Base, PEGASUS-Large, BART-large-CNN, Longformer-Encoder-Decoder (LED), and ProphetNet—using the original abstracts as reference summaries. Through rigorous comparative evaluation and domain-specific optimization, Longformer-LED emerges as the most effective model for summarizing complex mental health literature [78].

Other studies explore biases in multilingual depression classification using the RADAR-MDD dataset, which comprises English, Spanish, and Dutch samples. This study applies large language models (LLMs) to classify depression severity from speech-derived text, with an emphasis on biases related to language, age, and gender. We transcribe free-response speech recordings with Whisper and feed the resulting transcripts into several pre-trained LLMs—flanT5, RoBERTa, BERT, GPT-2, and mBERT—each augmented with a multilayer perceptron for binary classification of high versus low symptom severity. Models are evaluated on both balanced and unbalanced datasets stratified by language, gender, and age to assess performance disparities. Our key contribution is a systematic analysis of demographic biases in multilingual depression classification; we show that balancing for age has a more pronounced effect on model performance than balancing for gender, and that language differences significantly impact accuracy [76, 63].

Another work [80] investigates symptom-based depression severity estimation using only textual data from the DAIC-WOZ dataset, which comprises 189 structured clinical interviews. Instead of framing depression detection as a binary classification or regression task, this research estimates the severity of each symptom individually. The methodology compares both encoder-based and decoder-based large language models using in-context learning (ICL) strategies—including zero-shot, few-shot, and chain-of-thought (CoT) prompting—as well as parameter-efficient fine-tuning (PEFT) techniques such as LoRA. Models evaluated include ModernBERT, Mistral-7B, LLaMA-3, DeepSeek-R1, and proprietary architectures like Gemini-2.0-Flash. The key contribution demonstrates that zero-shot ICL configurations can outperform fine-tuned models and that reasoning-tuned variants such as DeepSeek-R1-8B achieve higher symptom-level accuracy [80].

4.2.2 Neuroimaging-Based Prediction

Longitudinal neuroimaging data can be used to predict the trajectory of depression or treatment response.

GNNs GNNs are widely applied to predict outcomes from neuroimaging data. Study [71] involves data collection from 79 participants, including 25 Android and 54 iPhone users (ages 18–25) at the University of Connecticut. Data types collected comprised weekly QUIDS survey results, clinical diagnoses obtained through standardized medical evaluations, Fitbit activity and sleep records, raw GPS trajectories, and categorical GPS data. Exploratory analysis utilized heatmap visualizations to examine spatial movement patterns. Graph Neural Networks (GNNs) were then applied by constructing participant-similarity graphs, where edges were defined using various distance metrics—Euclidean similarity for continuous features and Levenshtein edit distance for categorical GPS sequences. The efficacy of each metric was assessed via K-Nearest Neighbors and spectral clustering; final clustering outputs were compared against clinical labels using F1 scores to identify the optimal metric for graph construction [71]. A parallel study with the same cohort and data modalities (including auxiliary metadata) employed an identical methodology, reaffirming the comparative evaluation of distance metrics in GNN-based depression outcome prediction [44].

Another important study [77] integrates functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) data from the EMBARC clinical trial to predict treatment responses in patients with major depressive disorder (MDD). Pre-treatment resting-state fMRI and EEG scans were obtained from 130 patients randomized to sertraline and 135 randomized to placebo. Functional connectivity matrices for each modality were constructed and augmented via Common Orthogonal Basis Extraction (COBE). These matrices were then encoded by parallel graph neural networks with dynamic, learnable weight and scaling matrices to capture spatial dependencies and enable interpretability of region-specific contributions. Encoded feature embeddings were fused through a modality correlation maximization strategy to produce joint representations, which fed into a multilayer perceptron (MLP) for treatment-outcome prediction. The framework identified key predictive brain networks—such as the frontoparietal control network and limbic system—associated with differential antidepressant and placebo responses [77].

4.2.3 Multimodal Prediction

Combining multiple data streams over time offers a powerful approach for early and accurate prediction.

GNNs MentalNet is designed for early depression detection by integrating user interaction data and textual content from social media within a Deep Graph Convolutional Neural Network (DGCNN). The architecture consists of: i) ego-network feature extraction: user interactions (replies, mentions, quote-tweets) are encoded with an LSTM autoencoder to derive node features, ii) heterogeneous graph construction: nodes represent users and posts, with edges capturing interaction types, iii) graph convolution: a DGCNN processes the heterogeneous graph using doubly stochastic normalization to stabilize learning, and iv) classification: the final node embeddings are fed into a classifier to predict depression risk. By stabilizing graph learning and capturing rich interaction patterns, MentalNet achieves up to 19% improvements in precision, recall, and F1 score over existing methods, demonstrating its effectiveness for social media-based depression detection [67, 68].

LLMs & Transformers Large language models can integrate diverse multimodal inputs for predictive tasks. Study [69] introduces a framework that enhances mental health assessment by combining ensemble machine learning with LLMs. Using a dataset of 41,000 mental health entries, we compared AdaBoost, Voting, Bagging, and Random Forest models, identifying Random Forest as the most effective. The predictive workflow comprises: i) processing a user query with the Random Forest model to predict potential mental health issues, and ii) forwarding the prediction via an API to Google Gemini, which generates personalized insights based on the predicted condition [69]. Another study employs emotion prompts to guide LLMs in fusing heterogeneous signals for enhanced multimodal depression detection accuracy [75].

5 Datasets

The advancement of AI-driven depression detection depends critically on high-quality, well-annotated datasets. These resources are indispensable for training, validating, and benchmarking new computational models. Datasets in this field are diverse, reflecting the multifaceted nature of depression and spanning several data modalities. We categorize these resources into three broad types:

- **Clinical and Interview-Based Datasets:** Structured, multimodal data collected in controlled clinical or research settings, often including audio, video, and questionnaire responses.
- **Physiological and Neuroimaging Datasets:** Objective biological markers captured via modalities such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI).
- **Social Media Datasets:** Large-scale, unstructured text and metadata harvested from online platforms, reflecting real-world user behaviors and language.

This section introduces three of the most influential public datasets—each representative of one of the categories above—that have been instrumental in advancing research in AI-driven depression detection.

5.1 Distress Analysis Interview Corpus (DAIC-WOZ)

The Distress Analysis Interview Corpus (DAIC), particularly its Wizard-of-Oz (WOZ) subset, is a widely adopted benchmark in multimodal depression detection [81]. It comprises conversational interviews designed to facilitate the automatic identification of psychological distress indicators—such

as depression, anxiety, and PTSD. Each session pairs a human participant with Ellie, a virtual interviewer avatar controlled by a concealed human operator in a Wizard-of-Oz setup. This approach fosters naturalistic, empathetic dialogue, enabling dynamic, responsive interactions that elicit rich verbal and nonverbal cues associated with mental health status [81].

The DAIC-WOZ corpus comprises 189 interview sessions, divided into training (107), development (35), and test (47) sets. Sessions range from 7 to 33 minutes in length. Each session provides synchronized high-fidelity facial video, audio recordings, and time-aligned text transcriptions of the dialogue. Each participant’s data is annotated with a Patient Health Questionnaire (PHQ-8) score to quantify depression severity, with a score of 10 or higher indicating moderate depression. A binary depression label based on this threshold is also included. Additionally, the corpus offers pre-extracted feature sets—facial action units from video and acoustic features (e.g., pitch, intensity) from audio—enabling researchers to work without extensive signal-processing expertise. Owing to its structured design, clinical grounding, and rich multimodal features, DAIC-WOZ is an invaluable benchmark for developing and comparing models that fuse verbal and nonverbal cues. A sample of the transcribed data is shown in Table 4.

Table 4: An illustrative sample from the DAIC-WOZ dataset transcript [81].

Speaker	Utterance
Ellie (Agent)	So, tell me about yourself.
Participant 300	I’m from Los Angeles. I’m a student at U S C. I’m twenty one. I love to cook. uhm I don’t know what else.
Ellie (Agent)	Tell me about something you did recently that you really enjoyed.
Participant 300	I had some friends over for dinner last week. I made uhm homemade pasta and some salads and it was just a really lovely evening.

5.2 Multi-modal Open Dataset for Mental-disorder Analysis (MODMA)

The MODMA dataset was developed to address the need for open, high-quality physiological and behavioral data in mental health research, with a focus on Major Depressive Disorder (MDD) [82]. It comprises a comprehensive corpus that includes both behavioral recordings (audio and video) and direct neurophysiological signals, which are often absent from publicly available datasets. Data were collected from 53 participants—24 diagnosed with MDD and 29 healthy controls—recruited from a university and a psychiatric hospital. The experimental protocol included multiple sessions, featuring a resting-state period and a task-based period during which participants viewed emotionally evocative video clips designed to elicit varying affective responses.

The MODMA dataset comprises four primary modalities for each participant:

1. High-density 128-channel electroencephalography (EEG) recorded with a Biosemi ActiveTwo system, providing detailed information on brain dynamics.
2. Audio recordings from clinical interviews conducted using the Hamilton Depression Rating Scale (HAM-D).
3. Eye movement data captured with Tobii Pro Glasses 2.
4. Facial expression videos recorded throughout each session.

In addition to the MDD diagnosis, each record is annotated with scores from multiple clinical scales, including the Patient Health Questionnaire (PHQ-9), the Generalized Anxiety Disorder scale (GAD-7), and the Beck Depression Inventory (BDI-II). The inclusion of high-density EEG data makes MODMA an essential resource for researchers developing graph neural network models to analyze brain connectivity or hybrid models that fuse neurophysiological signals with behavioral cues [82]. An overview of the data collected per participant is shown in Table 5.

Table 5: Data modalities collected per participant in the MODMA dataset [82].

Data Type	Specification
EEG	128-channel Biosemi ActiveTwo system
Audio Interview	Recordings from clinical interviews (HAMD)
Eye Movement	Tobii Pro Glasses 2 eye tracker data
Facial Video	High-resolution video of facial expressions
Clinical Scores	PHQ-9, GAD-7, BDI-II, and clinical diagnosis

5.3 Reddit Self-reported Depression Diagnosis (RSDD) Dataset

Social media-derived datasets are essential for studying depression in naturalistic, real-world settings. The Reddit Self-Reported Depression Diagnosis (RSDD) dataset exemplifies this approach by enabling text-based depression detection with more reliable labels than keyword-based collection methods [83]. RSDD defines its “depressed” class strictly as users who explicitly self-report a formal depression diagnosis, thereby distinguishing clinical cases from transient expressions of sadness.

The RSDD dataset is constructed by identifying Reddit posts containing explicit indicators of a formal depression diagnosis (e.g., “I was diagnosed with depression,” “my doctor diagnosed me with depression”). For each such user, the entire posting history is retrieved to form the depressed cohort. A matched control group is then assembled by selecting users with similar activity levels—such as comparable post and comment counts—from a broad range of non-mental-health subreddits. The final corpus comprises the posting histories of 9,921 depressed users and 107,316 controls, providing a large-scale resource for training text-based models. Although this approach improves ecological validity and label quality over simple keyword searches, it is constrained by the inherent uncertainty of self-reported diagnoses. Nevertheless, RSDD remains invaluable for developing scalable models on extensive, unstructured, and longitudinal text data [83]. A sample is shown in Table 6.

Table 6: An illustrative example of posts from the RSDD dataset [83].

User Label	Post Snippet (from post history)
Depressed	”I was officially diagnosed with severe depression today. I don’t know how to feel. It’s a relief to have a name for it but now it feels so... real.”
Control	”Just finished building my new PC. The cable management was tough but it’s running Cyberpunk 2077 on ultra settings, so worth it!”

6 Evaluation Metrics

Rigorous assessment and comparison of AI-based depression detection models require standardized evaluation metrics. Metric selection depends on the model’s primary objective, which typically falls into one of two categories: classification or regression. Classification tasks assign discrete labels (e.g., depressed vs. non-depressed), while regression tasks predict continuous severity scores. In the following, we introduce the key metrics used in the depression detection literature for both tasks, present their mathematical definitions, and cite examples of their application.

6.1 Metrics for Classification Tasks

Classification models are most commonly evaluated using metrics derived from the confusion matrix, which tabulates the number of correct and incorrect predictions for each class. The core components of the matrix are True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy Accuracy measures the proportion of correct predictions among all samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Although easy to interpret, accuracy can be misleading on imbalanced datasets, since a model may achieve high scores by always predicting the majority class. In Zhu et al. [33], accuracy is employed as the primary metric for EEG-based depression recognition.

Precision, Recall, and F1-Score These metrics provide a more nuanced perspective on model performance, especially when the underlying class distribution is imbalanced.

- **Precision**, or positive predictive value, is the proportion of true positives among all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

It answers the question, “Of all individuals flagged as depressed by the model, how many were actually depressed?” High precision is vital in clinical settings to minimize false alarms. This metric is used in the fNIRS-based depression detection study [35].

- **Recall**, also known as sensitivity or true positive rate, is the proportion of actual positives correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

It answers the question, “Of all individuals who are truly depressed, how many did the model correctly identify?” High recall is vital in clinical contexts to ensure few cases are missed. This metric is also used in the fNIRS study [35].

- **F1-Score** is the harmonic mean of Precision and Recall, providing a single score that balances both metrics. It is particularly useful for evaluating models on imbalanced datasets where both

minimizing false positives and false negatives is important. The F1-score is a key evaluation metric in the multimodal study by [60].

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}.$$

Specificity, also known as the true negative rate, is the proportion of actual negatives correctly identified by the model. It complements recall by confirming the model can effectively identify healthy individuals. This metric is used in the DepressionGraph framework [47].

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Area Under the ROC Curve (AUC-ROC) measures a model’s discriminative ability independent of any specific threshold. The ROC curve plots the true positive rate (recall) against the false positive rate ($1 - \text{specificity}$) across all classification thresholds. The AUC represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative one. This metric is a primary evaluation criterion for the LGMF-GNN model [34].

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t) d(\text{FPR}(t)).$$

6.2 Metrics for Regression Tasks

For tasks that predict a continuous depression severity score, such as a PHQ-8 or BDI score, evaluation metrics measure the average error or agreement between the predicted (y') and true (y) scores for a set of n samples.

Mean Absolute Error (MAE) MAE measures the average absolute difference between the predicted and actual depression severity scores. It is easy to interpret since it represents the average prediction error in the original units of the score. MAE is used by [54] to evaluate the performance of LLMs in predicting PHQ-8 scores.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i|$$

Root Mean Square Error (RMSE) RMSE is the square root of the average of the squared differences between predicted and actual scores. By squaring the errors, it gives significantly more weight to larger errors, making it a useful metric when large deviations are particularly undesirable. RMSE is used to evaluate the SGP-SL model’s ability to predict PHQ-9 scores from EEG data [48].

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2}.$$

Concordance Correlation Coefficient (CCC) CCC measures the agreement between predicted and true scores by evaluating both their linear correlation and their deviation from the 45° line of perfect concordance. A CCC of 1 indicates perfect agreement, −1 perfect disagreement, and 0 no agreement. It is more robust than Pearson correlation for assessing prediction accuracy and is used as a key metric in the Chain-of-Thought LLM study [54].

$$\text{CCC} = \frac{2 \rho \sigma_y \sigma_{y'}}{\sigma_y^2 + \sigma_{y'}^2 + (\mu_y - \mu_{y'})^2},$$

where ρ is the Pearson correlation coefficient, and μ and σ^2 denote the means and variances of the respective variables.

7 Future Research Directions and Open Challenges

The body of work reviewed in this survey highlights significant progress in applying AI to depression detection. Analysis of the current literature reveals several prominent trends: the dominance of graph neural networks for neuroimaging; the rise of large language models for text and multimodal analysis; a strong push toward multimodal fusion; and an emerging focus on explainability and fairness. While these trends demonstrate the field’s maturation, they also underscore several open challenges and promising avenues for future research.

7.1 From Multimodal Correlation to Causal Fusion

A clear trend is the move toward sophisticated multimodal fusion, combining everything from EEG and audio signals [43] to fMRI, structural MRI, and health records [34]. Current models excel at identifying correlations between these data streams; however, the next critical step is to move from correlation to causation. Future research should prioritize developing causal inference models that untangle the complex interplay among modalities. For example, can changes in brain connectivity (from fMRI) be shown to causally influence specific vocal features (from audio) or linguistic patterns (from text)? Answering such questions would not only improve model robustness but also deepen insights into the neurobiological underpinnings of depression—a direction already explored in unimodal neuroimaging studies [59]. Furthermore, there is a need for lightweight fusion models that are computationally efficient enough for real-time deployment on personal devices such as smartphones, a crucial step toward continuous, real-world mental health monitoring [64].

7.2 Enhancing Model Interpretability for Clinical Trust

As models grow more sophisticated, there is an increasing emphasis on ensuring their explainability and interpretability in clinical settings. This focus directly counters the black-box nature of many deep learning systems. Recent innovative approaches include developing retrieval-augmented generation (RAG) frameworks that ground predictions in textual evidence from clinical transcripts [55], employing capsule networks to extract symptom-specific features aligned with the PHQ-9 [31], and designing GNNs capable of identifying clinically relevant brain subgraphs [38].

Despite this progress, future work must move beyond post-hoc explanations toward models that are inherently interpretable by design. Research should focus on developing systems that generate clear, evidence-based explanations in natural language, understandable to both clinicians

and patients. For example, a model could not only predict high depression risk but also state that its reasoning is based on “reduced speech rate, increased use of first-person pronouns, and anomalous connectivity in the frontoparietal control network.” Validating these explanations with mental health professionals will be crucial for building the trust necessary for clinical adoption.

7.3 Shifting from Diagnosis to Longitudinal Prediction and Intervention

The majority of reviewed studies focus on diagnosing current depression. While this remains crucial, the next frontier is forecasting depression onset, trajectory, and treatment response over time. Achieving this requires gathering and analyzing longitudinal data from diverse sources—such as wearable sensors and periodic self-assessments [72]. Although a handful of studies have begun predicting treatment response using neuroimaging [77], this domain remains largely unexplored. Ultimately, these predictive models should be integrated into personalized, just-in-time intervention systems that deliver timely, targeted support—whether suggesting coping strategies or prompting users to seek professional help at the earliest warning signs.

7.4 Addressing Data Scarcity, Privacy, and Diversity

The scarcity of large-scale, high-quality, and demographically diverse datasets remains a major bottleneck in the field, driven by the sensitive nature of mental health data and the high costs of clinical data collection. Future research must adopt privacy-preserving machine learning approaches. For instance, federated learning can enable model training on decentralized data from multiple hospitals or user devices without compromising patient confidentiality. Moreover, fairness studies have highlighted demographic imbalances in existing datasets [66]. Developing more sophisticated, clinically validated data augmentation techniques—especially for underrepresented groups—will be critical for building robust, equitable models that generalize effectively to the broader population.

7.5 Cross-lingual and Cross-cultural Generalization

Much of the current research, particularly LLM-based work, relies on English-language datasets such as DAIC-WOZ; however, studies on multilingual corpora show that performance can vary substantially across languages and cultural contexts [76]. Depressive symptom expression is not universal, and models trained on one cultural group may not generalize to others. Future research should prioritize culturally aware model development and validation. This necessitates curating diverse, multilingual datasets and designing flexible architectures—either by enabling efficient adaptation to new languages or by learning universal, language-agnostic representations of depression.

8 Methodology

To compile and structure the literature for this survey, we employed a systematic and multi-stage methodology. The process was designed to ensure a comprehensive, relevant, and high-quality selection of papers, which were managed and tracked using a Google Sheet.

8.1 Search Strategy

Our primary database for literature collection was Google Scholar. We conducted a series of targeted searches using combinations of keywords to identify pertinent studies at the intersection of artificial intelligence and mental health. The core search terms included "depression detection" and "MDD detection," which were paired with a range of technical keywords such as "machine learning," "AI," "GNN," "LLM," and "DL" (Deep Learning). This strategy yielded an initial pool of 61 papers that formed the basis for our subsequent screening and selection process.

8.2 Inclusion and Quality Assessment

Each of the 61 papers from the initial pool was subjected to a rigorous quality assessment to determine its suitability for inclusion in this survey. We developed a 10-point scoring system to evaluate each paper across three key criteria:

- **Relevance:** How directly the paper addressed the task of depression detection using AI.
- **Venue Quality:** The reputation and impact of the journal or conference where the paper was published.
- **Recency:** The publication date, with a preference for more recent work to reflect the current state of the field.

A strict filtering protocol was applied: any paper that scored below 5 out of 10 on even one of these criteria was excluded from further consideration. This comprehensive review process ensured that our final selection was both current and of high academic standing, ultimately resulting in the 55 papers that form the core of this survey.

8.3 Categorization

A key contribution of this survey is its unique hierarchical classification framework, designed to provide a structured and intuitive overview of the research landscape. After a thorough analysis, we identified three primary variant factors among the selected papers and used them to build our taxonomy. Each paper was assigned to a single primary category at each level to ensure a clear and non-redundant classification.

The categorization was performed in the following hierarchical order:

- **Level 1: Task Type.** The first and most fundamental division was based on the primary goal of the research. Papers were classified into two major groups:
 - *Diagnosis:* Works focused on identifying the current depressive state of an individual.
 - *Prediction:* Works focused on forecasting the future onset or risk of depression.
- **Level 2: Data Type.** Within each task, papers were further classified based on the data modality used, which was the second most significant variant factor. This level includes categories such as Text, Voice, Neuroimaging, and Multimodal data.
- **Level 3: Methodology and Model.** Finally, at the deepest level of the hierarchy, papers were grouped by the specific AI model or methodology employed. As this was the most variant factor, it led to the creation of distinct classes such as Graph Neural Networks (GNNs), Large Language Models (LLMs), and Hybrid Models.

This hierarchical categorization strategy allows for a nuanced and detailed analysis of the field, enabling readers to navigate the literature based on their specific interests in clinical tasks, data sources, or computational techniques.

9 Conclusion

This survey has provided a comprehensive and systematically structured overview of the application of Artificial Intelligence in the detection of depressive disorders. By reviewing 55 recent important studies, we have mapped the current landscape, highlighting the significant progress made in leveraging computational models to analyze a diverse range of data modalities. Our unique hierarchical taxonomy, which categorizes research first by clinical task (Diagnosis vs. Prediction), then by data type, and finally by model class, offers a novel framework for understanding the field’s key trends and identifying specific areas of innovation.

Our review confirms a clear trend towards the adoption of sophisticated deep learning architectures. Graph Neural Networks have become the standard for modeling the complex, structured data of neuroimaging, while Large Language Models and Transformers are revolutionizing the analysis of text and multimodal interview data. Furthermore, we identified a growing and critical focus on addressing the practical challenges of clinical implementation, with an increasing number of studies dedicated to enhancing model explainability and ensuring algorithmic fairness. Despite these advancements, significant challenges remain, including the need to move from correlational to causal models, address data scarcity through privacy-preserving techniques, and ensure models are culturally and linguistically generalizable.

By synthesizing the state of the art and outlining these open challenges, this paper serves as a valuable resource for researchers, clinicians, and engineers. It not only provides a detailed map of what has been accomplished but also offers a clear roadmap for future research directions. Continued interdisciplinary collaboration will be essential to overcoming the existing hurdles and fully realizing the potential of AI to create objective, accessible, and effective tools that can transform mental healthcare.

References

- [1] American Psychiatric Association, Diagnostic and statistical manual of mental disorders, 5th edition, text revision (DSM-5-TR), American Psychiatric Association Publishing, Washington, DC, 2022.
- [2] World Health Organization, Depressive disorder (depression), WHO Newsroom Accessed: July 19, 2025 (March 2023).
URL <https://www.who.int/news-room/fact-sheets/detail/depression>
- [3] G. S. Malhi, J. J. Mann, Depression, *The Lancet* 392 (10161) (2018) 2299–2312.
- [4] S. Evans-Lacko, M. Knapp, The economic impact of depression in high-income countries: a systematic review, *The Journal of Mental Health Policy and Economics* 19 (1) (2021) 37.
- [5] H. J. Lethus, D. Schouten, O. A. van den Heuvel, D. J. Veltman, R. van de Mheen, Artificial intelligence for mental health: a systematic review of opportunities, challenges, and future directions, *Translational Psychiatry* 13 (1) (2023) 344.

- [6] A. Deng, R. Guidotti, M. A. Al-garadi, A. Sarker, Artificial intelligence in mental health, *Artificial Intelligence in Medicine* (2024) 102761.
- [7] J. Cheng, H. Li, J. Zhang, D. Tao, A review of multimodal data fusion for mental disorder detection, *IEEE Transactions on Affective Computing* (2024).
- [8] J. W. Gichoya, I. Banerjee, H. Trivedi, M. Pugh, P. Film, C. Campos, et al., Ai recognition of patient race in medical imaging: a modelling study, *The Lancet Digital Health* 4 (6) (2022) e406–e414.
- [9] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W.-l. Li, Depression detection from social media: A survey, *IEEE Transactions on Affective Computing* 12 (4) (2017) 915–935.
- [10] F. Apicella, A. Fagiolini, C. Gesi, Machine learning-based classification of major depressive disorder using structural and functional neuroimaging: a systematic review, *Molecular Psychiatry* 28 (1) (2023) 262–274.
- [11] A. P. Association, What is depression?, Washington, DC: American Psychiatric Association (2018).
- [12] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.
URL <https://openreview.net/forum?id=SJU4ayYgl>
- [13] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, *IEEE transactions on neural networks and learning systems* 32 (1) (2020) 4–24.
- [14] M. H. Chehreghani, Half a decade of graph convolutional networks, *Nat. Mach. Intell.* 4 (3) (2022) 192–193. doi:10.1038/s42256-022-00466-8.
URL <https://doi.org/10.1038/s42256-022-00466-8>
- [15] M. Zohrabi, S. Saravani, M. H. Chehreghani, Centrality-based and similarity-based neighborhood extension in graph neural networks, *J. Supercomput.* 80 (16) (2024) 24638–24663. doi:10.1007/s11227-024-06336-x.
URL <https://doi.org/10.1007/s11227-024-06336-x>
- [16] Y. Mohamadi, M. H. Chehreghani, Strong transitivity relations and graph neural networks, *CoRR* abs/2401.01384 (2024). arXiv:2401.01384, doi:10.48550/ARXIV.2401.01384.
URL <https://doi.org/10.48550/arXiv.2401.01384>
- [17] F. Gholamzadeh Nasrabadi, A. Kashani, P. Zahedi, M. Haghiri Chehreghani, Content augmented graph neural networks, *ACM Trans.* WebJust Accepted (Oct. 2024). doi:10.1145/3700790.
URL <https://doi.org/10.1145/3700790>
- [18] C. I. Kanatsoulis, E. Choi, S. Jegelka, J. Leskovec, A. Ribeiro, Learning efficient positional encodings with graph neural networks, in: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025, OpenReview.net, 2025.
URL <https://openreview.net/forum?id=AWg2tkbyd0>

- [19] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Vol. 2, IEEE, 2005, pp. 729–734.
- [20] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
- [22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2018, pp. 4171–4186.
- [23] E. Niedermeyer, F. L. da Silva, Electroencephalography: basic principles, clinical applications, and related fields, Lippincott Williams & Wilkins, 2005.
- [24] M. Teplan, Fundamentals of eeg measurement, Measurement science review 2 (2) (2002) 1–11.
- [25] M. Al-Farras, A. Al-Wabil, H. Al-Negheimish, Z. Al-Hussain, W. Al-Shehri, F. Al-Sultan, A review of eeg-based diagnosis of psychiatric disorders, Sensors 23 (17) (2023) 7481.
- [26] M. Deshpande, V. Rao, Depression detection using emotion artificial intelligence, in: 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 858–862. doi:10.1109/ISS1.2017.8389299.
- [27] X. Huang, F. Wang, Y. Gao, et al., Depression recognition using voice-based pre-training model, Scientific Reports 14 (2024) 12734. doi:10.1038/s41598-024-63556-0. URL <https://doi.org/10.1038/s41598-024-63556-0>
- [28] T. T. Prama, M. S. Islam, M. M. Anwar, I. Jahan, Ai-enabled deep depression detection and evaluation informed by dsm-5-tr, IEEE Transactions on Computational Social Systems 11 (5) (2024) 6453–6465. doi:10.1109/TCSS.2024.3382139.
- [29] M. O. Nusrat, W. Shahzad, S. A. Jamal, Multi class depression detection through tweets using artificial intelligence (2024). arXiv:2404.13104. URL <https://arxiv.org/abs/2404.13104>
- [30] S. Patil, N. Jagtap, K. Jadhav, A. Desai, A. Shaikh, Depression detection using convolutional neural network, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Accessed: Apr. 12, 2025 (2023).
- [31] H. Liu, C. Li, X. Zhang, F. Zhang, W. Wang, F. Ma, H. Chen, H. Yu, X. Zhang, Depression detection via capsule networks with contrastive learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 22231–22239. doi:10.1609/aaai.v38i20.30228.
- [32] V. Tejaswini, K. Sathya Babu, B. Sahoo, Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23 (1) (Jan. 2024). doi:10.1145/3569580. URL <https://doi.org/10.1145/3569580>

- [33] J. Zhu, C. Jiang, J. Chen, X. Lin, R. Yu, X. Li, B. Hu, Eeg based depression recognition using improved graph convolutional neural network, *Computers in Biology and Medicine* 148 (2022) 105815. doi:<https://doi.org/10.1016/j.combiomed.2022.105815>. URL <https://www.sciencedirect.com/science/article/pii/S0010482522005765>
- [34] S. Liu, J. Zhou, X. Zhu, Y. Zhang, X. Zhou, S. Zhang, Z. Yang, Z. Wang, R. Wang, Y. Yuan, X. Fang, X. Chen, Y. Wang, L. Zhang, G. Wang, C. Jin, An objective quantitative diagnosis of depression using a local-to-global multimodal fusion graph neural network, *Patterns* 5 (12) (2024) 101081. doi:<https://doi.org/10.1016/j.patter.2024.101081>. URL <https://www.sciencedirect.com/science/article/pii/S266638992400240X>
- [35] Q. Yu, R. Wang, J. Liu, L. Hu, M. Chen, Z. Liu, Gnn-based depression recognition using spatio-temporal information: A fnirs study, *IEEE Journal of Biomedical and Health Informatics* 26 (10) (2022) 4925–4935. doi:[10.1109/JBHI.2022.3195066](https://doi.org/10.1109/JBHI.2022.3195066).
- [36] Z. Chen, J. Deng, J. Zhou, J. Wu, T. Qian, M. Huang, Depression detection in clinical interviews with LLM-empowered structural element graph, in: K. Duh, H. Gomez, S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 8181–8194. doi:[10.18653/v1/2024.naacl-long.452](https://doi.org/10.18653/v1/2024.naacl-long.452). URL <https://aclanthology.org/2024.naacl-long.452/>
- [37] X. Zhang, H. Liu, K. Xu, Q. Zhang, D. Liu, B. Ahmed, J. Epps, When LLMs meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 146–158. doi:[10.18653/v1/2024.emnlp-main.8](https://doi.org/10.18653/v1/2024.emnlp-main.8). URL <https://aclanthology.org/2024.emnlp-main.8/>
- [38] K. Zheng, S. Yu, B. Li, R. Jenssen, B. Chen, Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck, *IEEE Transactions on Neural Networks and Learning Systems* (2024) 1–14doi:[10.1109/TNNLS.2024.3449419](https://doi.org/10.1109/TNNLS.2024.3449419).
- [39] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, A. Cohan, Improving the generalizability of depression detection by leveraging clinical questionnaires (2022). [arXiv:2204.10432](https://arxiv.org/abs/2204.10432). URL <https://arxiv.org/abs/2204.10432>
- [40] M. Danner, B. Hadzic, S. Gerhardt, S. Ludwig, I. Uslu, P. Shao, T. Weber, Y. Shiban, M. Ratsch, Advancing mental health diagnostics: Gpt-based method for depression detection, in: *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, 2023, pp. 1290–1296. doi:[10.23919/SICE59929.2023.10354236](https://doi.org/10.23919/SICE59929.2023.10354236).
- [41] T. Zhao, G. Zhang, Enhancing major depressive disorder diagnosis with dynamic-static fusion graph neural networks, *IEEE Journal of Biomedical and Health Informatics* 28 (8) (2024) 4701–4710. doi:[10.1109/JBHI.2024.3395611](https://doi.org/10.1109/JBHI.2024.3395611).
- [42] M. Sadeghi, B. Egger, R. Agahi, R. Richer, K. Capito, L. H. Rupp, L. Schindler-Gmelch, M. Berking, B. M. Eskofier, Exploring the capabilities of a language model-only approach for

- depression detection in text data, in: 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 2023, pp. 1–5. doi:10.1109/BHI58575.2023.10313367.
- [43] T. Chen, R. Hong, Y. Guo, S. Hao, B. Hu, Ms²-gnn: Exploring gnn-based multimodal fusion network for depression detection, IEEE Transactions on Cybernetics 53 (12) (2023) 7749–7759. doi:10.1109/TCYB.2022.3197127.
 - [44] J. Kim, M. Sohn, Graph representation learning-based early depression detection framework in smart home environments, Sensors 22 (4) (2022). doi:10.3390/s22041545.
URL <https://www.mdpi.com/1424-8220/22/4/1545>
 - [45] M. Zhang, D. Long, Z. Chen, C. Fang, Y. Li, P. Huang, F. Chen, H. Sun, Multi-view graph network learning framework for identification of major depressive disorder, Computers in Biology and Medicine 166 (2023) 107478. doi:<https://doi.org/10.1016/j.compbimed.2023.107478>.
URL <https://www.sciencedirect.com/science/article/pii/S0010482523009435>
 - [46] M. Li, X. Sun, M. Wang, Detecting depression with heterogeneous graph neural network in clinical interview transcript, IEEE Transactions on Computational Social Systems 11 (1) (2024) 1315–1324. doi:10.1109/TCSS.2023.3263056.
 - [47] Z. Xia, Y. Fan, K. Li, Y. Wang, L. Huang, F. Zhou, Depressiongraph: A two-channel graph neural network for the diagnosis of major depressive disorders using rs-fmri, Electronics 12 (24) (2023). doi:10.3390/electronics12245040.
URL <https://www.mdpi.com/2079-9292/12/24/5040>
 - [48] T. Chen, Y. Guo, S. Hao, R. Hong, Exploring self-attention graph pooling with eeg-based topological structure and soft label for depression detection, IEEE Transactions on Affective Computing 13 (4) (2022) 2106–2118. doi:10.1109/TAFFC.2022.3210958.
 - [49] F. Noman, C.-M. Ting, H. Kang, R. C.-W. Phan, H. Ombao, Graph autoencoders for embedding learning in brain networks and major depressive disorder identification, IEEE Journal of Biomedical and Health Informatics 28 (3) (2024) 1644–1655. doi:10.1109/JBHI.2024.3351177.
 - [50] S. Venkatapathy, M. Votinov, L. Wagels, S. Kim, M. Lee, U. Habel, I.-H. Ra, H.-G. Jo, Ensemble graph neural network model for classification of major depressive disorder using whole-brain functional connectivity, Frontiers in Psychiatry Volume 14 - 2023 (2023). doi:10.3389/fpsyt.2023.1125339.
URL <https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1125339>
 - [51] W. Ma, Y. Wang, N. Ma, Y. Ding, Diagnosis of major depressive disorder using a novel interpretable gcn model based on resting state fmri, Neuroscience 566 (2025) 124–131. doi:10.1016/j.neuroscience.2024.12.045.
URL <https://www.sciencedirect.com/science/article/pii/S0306452224007528>
 - [52] Y. Wang, W. Zheng, Y. Li, H. Yang, A hybrid graph neural network for enhanced eeg-based depression detection (2024). arXiv:2410.18103.
URL <https://arxiv.org/abs/2410.18103>

- [53] C. Sun, M. Jiang, L. Gao, Y. Xin, Y. Dong, A novel study for depression detecting using audio signals based on graph neural network, *Biomedical Signal Processing and Control* 88 (2024) 105675. doi:<https://doi.org/10.1016/j.bspc.2023.105675>.
URL <https://www.sciencedirect.com/science/article/pii/S1746809423011084>
- [54] S. Teng, J. Liu, R. K. Jain, S. Chai, R. Hou, T. Tateyama, L. Lin, Y. wei Chen, Enhancing depression detection with chain-of-thought prompting: From emotion to reasoning using large language models (2025). *arXiv:2502.05879*.
URL <https://arxiv.org/abs/2502.05879>
- [55] L. Zhang, Z. Gao, D. Zhou, Y. He, Explainable depression detection in clinical interviews with personalized retrieval-augmented generation (2025). *arXiv:2503.01315*.
URL <https://arxiv.org/abs/2503.01315>
- [56] X. Zhang, H. Liu, Q. Zhang, B. Ahmed, J. Epps, Speechrag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information (2025). *arXiv:2502.10950*.
URL <https://arxiv.org/abs/2502.10950>
- [57] Z. Xu, C. L. P. Chen, T. Zhang, Tflagl: A novel agent graph learning method using time-frequency eeg for major depressive disorder detection, *IEEE Transactions on Affective Computing* (2025) 1–14doi:10.1109/TAFFC.2025.3527459.
- [58] C. Yang, X. Dong, X. Zong, Frequency feature fusion graph network for depression diagnosis via fnirs (2025). *arXiv:2504.21064*.
URL <https://arxiv.org/abs/2504.21064>
- [59] S. Kim, S. H. Bong, S. Yun, D. Kim, J. H. Yoo, K. S. Choi, H. Park, H. J. Jeon, J.-H. Kim, J. H. Jang, B. Jeong, Neurobiologically interpretable causal connectome for predicting young adult depression: A graph neural network study, *Journal of Affective Disorders* 377 (2025) 225–234. doi:10.1016/j.jad.2025.02.076.
URL <https://www.sciencedirect.com/science/article/pii/S0165032725002824>
- [60] J. Cha, S. Kim, D. Kim, E. Park, Mogam: A multimodal object-oriented graph attention model for depression detection (2024). *arXiv:2403.15485*.
URL <https://arxiv.org/abs/2403.15485>
- [61] Z. Yan, F. Peng, D. Zhang, Decen: A deep learning model enhanced by depressive emotions for depression detection from social media content, *Decision Support Systems* 191 (2025) 114421. doi:10.1016/j.dss.2025.114421.
URL <https://www.sciencedirect.com/science/article/pii/S0167923625000223>
- [62] F. Zhu, J. Zhang, R. Dang, B. Hu, Q. Wang, MTNet: Multimodal transformer network for mild depression detection through fusion of EEG and eye tracking, *Biomedical Signal Processing and Control* 100 (2025) 106996. doi:10.1016/j.bspc.2024.106996.
URL <https://www.sciencedirect.com/science/article/pii/S1746809424010541>
- [63] U. Naseem, A. Dunn, J. Kim, M. Khushi, Detection of depression severity in social media text using transformer-based models, *Information* 16 (2) (2025) 114. doi:10.3390/info16020114.
URL <https://www.mdpi.com/2078-2489/16/2/114>

- [64] E. Lim, M. Jhon, J.-W. Kim, S.-H. Kim, S. Kim, H.-J. Yang, A lightweight approach based on cross-modality for depression detection, *Computers in Biology and Medicine* 186 (2025) 109618. doi:10.1016/j.combiomed.2024.109618.
URL <https://www.sciencedirect.com/science/article/pii/S0010482524017037>
- [65] J. Cheong, A. Bangar, S. Kalkan, H. Gunes, U-fair: Uncertainty-based multimodal multitask learning for fairer depression detection (2025). *arXiv:2501.09687*.
URL <https://arxiv.org/abs/2501.09687>
- [66] A. M. H. Kwok, J. Cheong, S. Kalkan, H. Gunes, Machine learning fairness for depression detection using eeg data (2025). *arXiv:2501.18192*.
URL <https://arxiv.org/abs/2501.18192>
- [67] T. Xing, Y. Dou, X. Chen, et al., An adaptive multi-graph neural network with multimodal feature fusion learning for mdd detection, *Scientific Reports* 14 (2024) 28400. doi:10.1038/s41598-024-79981-0.
- [68] A.-T. Kuo, H. Chen, Y.-H. Kuo, W.-S. Ku, Dynamic graph representation learning for depression screening with transformer (2023). *arXiv:2305.06447*.
URL <https://arxiv.org/abs/2305.06447>
- [69] S. Kamoji, S. Rozario, S. Almeida, S. Patil, S. Patankar, H. Pendhari, Mental health prediction using machine learning models and large language model, in: 2024 Second International Conference on Inventive Computing and Informatics (ICICI), 2024, pp. 185–190. doi:10.1109/ICICI62254.2024.00040.
- [70] K. Sabaneh, M. A. Salameh, F. Khaleel, M. M. Herzallah, J. Y. Natsheh, M. Maree, Early risk prediction of depression based on social media posts in arabic, in: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), 2023, pp. 595–602. doi:10.1109/ICTAI59109.2023.00094.
- [71] G. Luo, H. Rao, P. An, Y. Li, R. Hong, W. Chen, S. Chen, Exploring adaptive graph topologies and temporal graph networks for eeg-based depression detection, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023) 3947–3957. doi:10.1109/TNSRE.2023.3320693.
- [72] P. Bidja, Depressiongnn: Depression prediction using graph neural network on smartphone and wearable sensors, Honors scholar thesis, University of Connecticut (May 2019).
URL https://digitalcommons.lib.uconn.edu/srhonors_theses/607/
- [73] W. Qin, Z. Chen, L. Wang, Y. Lan, W. Ren, R. Hong, Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media (2023). *arXiv:2305.05138*.
URL <https://arxiv.org/abs/2305.05138>
- [74] S. M. Shah, S. A. Gillani, M. S. A. Baig, M. A. Saleem, M. H. Siddiqui, Advancing depression detection on social media platforms through fine-tuned large language models, *Online Social Networks and Media* 46 (2025) 100311. doi:10.1016/j.osnem.2025.100311.
URL <https://www.sciencedirect.com/science/article/pii/S2468696425000126>

- [75] S. Teng, J. Liu, H. Sun, S. Chai, T. Tateyama, L. Lin, Y.-W. Chen, Enhanced multi-modal depression detection with emotion prompts, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10889035.
- [76] P. A. Pérez-Toro, J. Dineley, R. Iniesta, et al., Exploring biases related to the use of large language models in a multilingual depression corpus, Research SquarePreprint (jan 2025). doi:10.21203/rs.3.rs-5731243/v1.
- [77] Y. Jiao, K. Zhao, X. Wei, et al., Deep graph learning of multimodal brain networks defines treatment-predictive signatures in major depression, Molecular Psychiatry (2025). doi:10.1038/s41380-025-02974-6.
- [78] A. Kumar, A. Sharma, S. R. Sangwan, Transformer-based abstractive summarization for depression detection literature for enhanced medical insights, AuthoreaPreprint (January 09 2025).
- [79] Y. Ibrahimov, T. Anwar, T. Yuan, Depressionx: Knowledge infused residual attention for explainable depression severity assessment (2025). arXiv:2501.14985. URL <https://arxiv.org/abs/2501.14985>
- [80] D. E. Merzougui, G. Dias, J. Pantin, F. Maurel, Evaluating large language models for depression symptom estimation, in: R. Bellazzi, J. M. Juarez Herrero, L. Sacchi, B. Zupan (Eds.), Artificial Intelligence in Medicine. AIME 2025, Vol. 15735 of Lecture Notes in Computer Science, Springer, Cham, 2025, pp. 589–599. doi:10.1007/978-3-031-95841-0_51. URL https://doi.org/10.1007/978-3-031-95841-0_51
- [81] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., The distress analysis interview corpus of human and computer interviews, in: Proceedings of the ninth international conference on language resources and evaluation (lrec'14), 2014, pp. 3123–3128.
- [82] Y.-W. Cao, Z.-W. Wei, Y.-S. Li, Y. Li, Z.-X. Qiu, Y.-X. Sun, J.-Y. Zhang, Q.-Q. Sun, D.-J. Liu, G.-Z. Chen, et al., Modma: A multi-modal open dataset for mental-disorder analysis, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1514–1522.
- [83] N. Schradin, A. Benta, T. Mitra, On the difficulties of specifically detecting depressive language on social media, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 15, 2021, pp. 577–588.