Artificial Intelligence for Depression Detection in Adults: A Systematic Review of Diagnostic Performance, Features, and Limitations

Adam Kirk

Submitted to: JMIR Mental Health
on: September 25, 2025

Table of Contents

Artificial Intelligence for Depression Detection in Adults: A Systematic Review of Diagnostic Performance, Features, and Limitations

Adam Kirk[1] MAOL

[1] Samford University Birmingham US

Corresponding Author:
Adam Kirk MAOL

Samford University
800 Lakeshore Drive
Birmingham
US

Abstract

Background: Mental health conditions, particularly major depressive disorder (MDD), are a significant and growing global concern. The World Health Organization (WHO) estimates that more than 264 million people worldwide suffer from depression, which is a leading cause of disability and premature mortality. In the United States alone, approximately 18.8 million adults experienced a major depressive episode in 2020. Despite the high prevalence, traditional diagnostic methods—such as clinical interviews and questionnaires—are often subjective, leading to inconsistent diagnoses and a high rate of missed cases. These limitations highlight an urgent need for more objective, efficient, and accessible screening tools to enable earlier detection and intervention.

Objective: This systematic review aims to synthesize recent evidence to answer the primary research question: "How effective are AI-based screening tools in identifying and diagnosing depression in adults, and what are their key features and limitations?" By thoroughly examining current research, this study seeks to highlight the successes of AI in mental health, identify common patterns, and pinpoint existing gaps to guide future research and improve clinical care.

Methods: This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. The search strategy involved querying several databases—PubMed, CINAHL, and ScienceDirect—for original research studies published between January 1, 2022, and June 30, 2025. Eligibility criteria required studies to focus on AI tools for depression screening or diagnosis in adults (aged 18 or older), provide measurable outcomes compared to a trusted standard, and use real-world data.

The selection process was conducted in two stages using Rayyan QCRI. First, 657 unique studies were screened by title and abstract, which narrowed the selection to 44. Second, a full-text review was conducted, resulting in the final inclusion of 21 studies. Data from these studies were extracted using a standardized form to collect information on study design, AI tool details, participant demographics, performance metrics (e.g., accuracy, sensitivity, specificity), and reported limitations. The quality and risk of bias for each study were assessed using the Joanna Briggs Institute (JBI) Critical Appraisal Tools.

Results: The 21 included studies demonstrated that AI tools have a significant capacity to detect and diagnose depression, often showing high performance metrics comparable to or exceeding traditional methods. The AI tools were remarkably diverse, leveraging various data modalities, including speech patterns, text from social media or health records, and physical/behavioral data (such as gait, brain activity, and sleep patterns). Performance metrics varied by tool and data type, but many studies reported high accuracy, with several reaching into the 80% and 90% ranges. For example, a multimodal tool combining text and speech achieved an accuracy of 99.81%, while a tool using polysomnography sleep data reached 96.88% accuracy. The diagnostic performance of these tools was consistently compared against trusted standards like the PHQ-9, BDI-II, or clinical diagnoses based on DSM-5 criteria.

While the results were promising, common limitations were identified across studies. Many focused on specific, limited cohorts (e.g., college students or older adults), which limits the generalizability of the findings. The reliance on cross-sectional designs also prevented insights into long-term tracking of depression. Furthermore, most of the AI tools had not yet been tested in real-world clinical settings. The review also noted that some complex AI models, like large language models, may struggle to interpret subtle human cognitive constructs crucial for accurate diagnosis.

Conclusions: This systematic review confirms the significant promise of AI-based tools for enhancing the early detection of depression in adults. These tools offer an objective, scalable, and accessible alternative to traditional methods, potentially easing the burden on both patients and clinicians. By leveraging diverse data modalities from everyday activities, AI can make mental health screening less intrusive and more efficient. However, to fulfill this potential, future research must address key challenges, including the need for more diverse and larger datasets, the use of longitudinal study designs, and robust real-world clinical validation. By focusing on these areas, researchers can help unlock AI's full capacity to transform mental health care, ensuring more people receive timely and effective support. Clinical Trial: N/A

(JMIR Preprints 25/09/2025:84862)
DOI: https://doi.org/10.2196/preprints.84862

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Artificial Intelligence for Depression Detection in Adults: A Systematic Review of Diagnostic Performance, Features, and Limitations

Introduction

Mental health issues, especially major depressive disorder (MDD), are a growing concern worldwide. They impact millions and cause significant personal, social, and economic burdens. The World Health Organization (WHO) estimates that over 264 million people live with depression, making it a leading cause of disability and a contributor to early mortality (World Health Organization, 2020). In the United States, around 18.8 million adults experienced a major depressive episode in 2020 (Villarroel & Terlizzi, 2022). These numbers highlight the urgent need for better, more accessible methods to identify and address depression early.

Traditional diagnostic methods, like clinical interviews and questionnaires, rely on subjective judgment and self-reporting, leading to inconsistencies. These approaches can miss cases or delay treatment, with studies suggesting up to half of people with depression in primary care settings go undiagnosed (Mitchell et al., 2011). Such gaps underscore the need for innovative tools to facilitate earlier detection and support.

Recent advancements in Artificial Intelligence (AI), including machine learning and deep learning, offer promising possibilities in mental healthcare. AI tools can process vast amounts of data quickly and effectively, detecting patterns and improving measurement (Graham et al., 2019; Topol, 2019). They can analyze diverse data—such as speech patterns (Cummins et al., 2015), text from social media (Shin et al., 2024), physiological signals from wearable devices (Abd-alrazaq et al., 2023), gait (walking patterns) (Zhao et al., 2021), brain imaging (Liu et al., 2024), and chatbot conversations (Torous et al., 2021) —to identify signs of depression more accurately. These technologies offer hope for more objective, efficient, and accessible mental health screening, overcoming limitations of traditional methods.

While AI holds great potential for mental health screening, its efficacy, features, and limitations, especially across diverse adult populations, require thorough understanding. This systematic review aims to synthesize current research to answer: "How effective are AI-based screening tools in identifying and diagnosing depression in adults, and what are their key features and limitations?". By carefully reviewing the evidence, this study will highlight successes, identify patterns, and pinpoint gaps to guide future research and improve mental health care

Methods

This systematic review investigates how well Artificial Intelligence (AI)-based tools identify and diagnose depression in adults, examining their strengths, features, and limitations. I followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines to ensure clarity and thoroughness (Page et al., 2021).

Eligibility Criteria

The review included studies that met the following criteria: participants were adults aged 18 or older; the tool studied was an AI-based tool using machine learning or deep learning for depression screening or diagnosis; results shared measurable outcomes like accuracy, sensitivity, specificity, and AUC compared to trusted methods (e.g., clinical diagnoses, PHQ-9, BDI-II, GDS-15, GAD-7); study types were original research (e.g., randomized trials, cohort studies); and the study included real-world data to test the AI tool. The focus was on initial screening or diagnosis, not treatment tracking. Studies had a minimum sample size of 10 participants, were available in full English text, and were published between January 1, 2022, and June 30, 2025.

Studies were excluded if they were not original research (e.g., reviews, opinion pieces), did not focus on AI tools for depression screening or diagnosis, lacked measurable results, studied non-adult ages, focused on treatment monitoring, were published outside the specified date range, or were not in English.

Search Strategy

To find relevant studies, several databases were searched for articles published between January 1, 2022, and June 30, 2025. The databases included PubMed, CINAHL, and ScienceDirect. Search terms were crafted to find studies about AI, depression screening, and adults, with terms like "artificial intelligence," "AI," "diagnos*," "depression," and "adult*." All searches were completed on July 1, 2025. The searches returned 284 records from PubMed, 314 from CINAHL, and 72 from ScienceDirect.

Study Selection

All records were imported into Rayyan QCRI. After removing 13 duplicates, 657 unique studies remained. The selection process involved two steps. First, a title and abstract review was conducted on the 657 records, excluding 613 that did not fit the criteria, leaving 44 studies for full-text review. Second, the full texts of the 44 studies were read and eligibility criteria were applied. This step excluded 23 studies for reasons such as unsuitable study design (n=15), irrelevant outcomes (n=6), incorrect publication type (n=1), or non-adult participants (n=1). This process resulted in 21 studies being included for analysis.

Data Extraction

A customized standardized form was used to collect key information from the 21 included studies. The collected data included study type, AI tool details (name, methods, input types, data sources), participant information (number, age, gender, race/ethnicity, location), performance results (sensitivity, specificity, AUC, accuracy), depression assessment tools used, and study limitations and suggestions for future research.

Quality Assessment

I evaluated the quality and risk of bias in each study using the Joanna Briggs Institute (JBI) Critical Appraisal Tools for observational and diagnostic studies (Joanna Briggs Institute, 2017). Studies were rated as low, moderate, or high risk of bias.

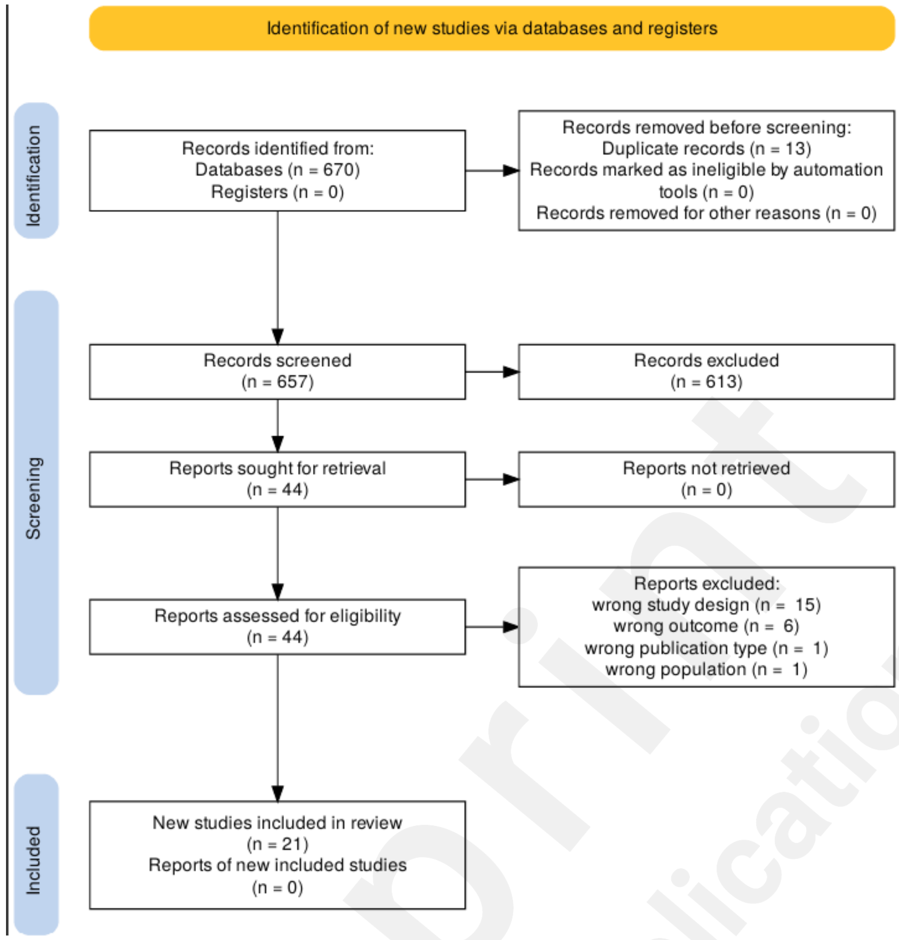
Data Synthesis

Since studies varied in AI methods, participants, and results, I summarized findings in a narrative format, using tables and text to describe effectiveness, tool features, and limitations. I reported ranges and averages for performance metrics but did not statistically combine data due to expected differences.

Results

Study Selection

My search yielded 670 records. After deduplication, 657 unique studies remained. Initial title and abstract screening excluded 613 records. A full-text review of the remaining 44 studies led to 23 more exclusions. Ultimately, 21 studies were included for analysis. Figure 1 (PRISMA flow diagram) details this process.



Details of Included Studies

The 21 included studies explored AI's role in depression detection in adults, employing various designs such as cross-sectional, model-building, and observational studies. The AI tools themselves demonstrated remarkable diversity, using a wide range of data inputs. Some were **speech-based**, analyzing acoustic features from free-form or prompted speech, as seen in studies using tools like Kintsugi Voice or wav2vec 2.0. Others were **text-based**, with Large Language Models (LLMs) like GPT-3.5 or ChatGPT-4, or natural language processing (NLP) models analyzing diary entries, social media posts, or patient messages. A third category used **physical and behavioral data**, examining gait from wearable sensors, brain activity from EEG or fMRI, sleep patterns from polysomnography, or facial expressions. Several studies also employed a **multimodal** approach, combining different data types such as speech and brain scans or text and gait. Finally, some tools were **record-based**, utilizing electronic health records (EHR) or national health surveys (NHANES). The participant demographics across these studies were varied, spanning young adults to older adults and representing countries like the United States, China, Canada, Brazil, Japan, and Korea. Most studies included both men and women, with specific populations including graduate students, Chinese university students, cancer patients, and diabetic patients. Depression was typically measured using validated tools like the PHQ-8/9, BDI-II, GDS-15, GAD-7, or clinical diagnoses based on DSM-5 criteria. A comprehensive overview of each study's design, AI type, population, and assessment method is provided in Table 1.

Table 1: Characteristics of Included Studies

	Study Design	AI Tool Used (What Kind of AI; What Data it Analyzed)	Participant Demographics	Depression Assessment Method
Johnson et al., 2025	Direct research	GPT (Large Language Model); Analyzed text (how people explain causes for things)	930 college students; Average age 18.75 (17-22 years); 60% women, 40% men	DAS, BDI-II questionnaires
Bastos et al., 2025	Direct research (looking at relationships)	Machine Learning; Analyzed self-reported academic stress	172 graduate students; Average age 28.0 (21-50 years); 67.4% women	PHQ-9 questionnaire
Wei et al., 2025	Direct research	Machine Learning (like AdaBoost, SVM); Analyzed speech sounds	332 university students in China; Average age 20.3 (18-24 years); 50.3% women	BDI-II questionnaire
Liu et al., 2024	Direct research (developing/testing a model)	LGMF-GNN (Deep Learning); Analyzed brain scans (fMRI, sMRI) and health records	2442 people (1260 with depression, 1182 healthy); Average age 29-43 years; From China, Japan, Russia	Psychiatrist diagnoses (using DSM-IV/DSM-5), MINI/SCID interviews
Marieke M. van Buchem et al., 2024	Direct research (developing/validating a tool)	NLP (Natural Language Processing); Analyzed patient messages	3312 cancer patients; Average age 61 years; 60% women; From Stanford Cancer Center, USA	Clinical diagnosis (doctor's diagnosis, prescriptions, or referrals)
Maekawa et al., 2024	Direct research	Bayesian Networks; Analyzed national health surveys and community data	Large groups from national datasets; Ages 15-75 (one group), 18+ (another); From Brazil, Chile	PHQ-9 questionnaire (score 10 or higher)
Byeon, 2023	Direct research	Machine Learning (like LASSO, RF, XGBoost); Analyzed past electronic health records	33,524 diabetic patients in Korea; Average age 55.4 years; 53.6% women	PHQ-8 questionnaire (score 10 or higher)
Hong et al., 2025	Direct research	TIMEX-D (Deep Learning); Analyzed text and sounds	Used public datasets (DAIC-WOZ/EDAIC); Average age 36.64 years (for DAIC-WOZ)	PHQ-8 questionnaire
Huang et al., 2024	Direct research	wav2vec 2.0 (Deep Learning); Analyzed voice data	Used a public dataset (DAIC-WOZ); Average age 36.64 years	PHQ-8 questionnaire
Ding et al.,	Direct research	PM-SED (Deep Learning); Analyzed emotional	187 participants in China; Average	PHQ-9, BDI-II

2025		speech	age 20.89 years; 59 men, 128 women	questionnaires, and clinician assessment (HAM-D)
Liu et al., 2025	Direct research (data collected at one time)	ChatGPT-4 (Large Language Model); Analyzed responses to a structured interview	200 college students; Likely 18+ years old	PHQ-9, GAD-7 questionnaires
Mazur et al., 2025	Direct research (data collected at one time)	Kintsugi Voice (Machine Learning); Analyzed voice patterns from free speech	14,898 adults (18-93 years, average 37.3); 27.3% men; From US, Canada	PHQ-9 questionnaire (score 10 or higher)
Lee et al., 2024	Direct research (developing/validating a model)	Deep Learning (CNN, LSTM, ResNet); Analyzed activity and sleep data from wearables	352 participants (60-90 years, average 72.48); 73.01% women	Clinical diagnosis
Jia et al., 2025	Direct research	MHA-GCN_ViT (Deep Learning); Analyzed brain waves (EEG) and speech signals	53 participants (24 with depression, 29 healthy); Ages 16-56 years; From China	Clinical diagnosis (from MODMA dataset)
Liu et al., 2025	Direct research (developing/validating a model)	Machine Learning/Deep Learning (Text-guided); Analyzed walking patterns and text from questionnaires	200 participants from D-Gait dataset in China; Ages 17-25 years	Combined SDS, PHQ-9, GAD-7 scores for confirmed diagnosis
Thakre et al., 2022	Direct research (looking back at old data)	Deep Learning (Xception); Analyzed sleep study data (transformed to images)	940 adults (18-90 years, average 42.7); 51.5% men, 48.5% women; From a sleep center in USA	Clinical diagnoses
Liu et al., 2023	Direct research (developing/validating/testing a model)	PRA-Net (Deep Learning); Analyzed facial expressions	Used public datasets (AVEC2013: 300, AVEC2014: 100 adults)	BDI-II scores
Vu et al., 2025	Direct research	Machine Learning (like LR, RF, XGBoost); Analyzed NHANES survey data	Used NHANES data (adults in USA)	PHQ-9 questionnaire (score 10 or higher)
Susanty et al., 2023	Direct research	Machine Learning (like XGBoost, RF); Analyzed walking patterns from wearables	330 participants in China; Average age 70.8 years; 55.4% women	GDS questionnaire (score 5 or higher) and medical doctor/records review
Georgescu et al., 2024	Direct research (observational)	Machine Learning (LR, XGBoost); Analyzed vocal features during a memory task	137 participants in Brazil (60-91 years, average 72.3); 60.6% women	GDS-15, GAD-7 questionnaires
Shin et al., 2024	Direct research (validating a new AI measurement tool)	Large Language Models (GPT-3.5); Analyzed user-written diary text	100 adult participants (average age 26.6, 19+ years); 53% women, 47% men	Psychiatrist diagnosis (DSM-5) and PHQ-9 questionnaire

Diagnostic Performance Metrics

The 21 studies consistently demonstrated AI tools' significant capacity to detect depression, with performance contingent on the AI approach, data type, and studied population. Accuracy varied from approximately 59.85% (Liu et al., 2025) to 99.81% (for a multimodal tool combining text and speech) (Hong et al., 2025). Many studies achieved accuracy in the 80s and 90s, indicating promising results. For instance, a study using polysomnography sleep data reached 96.88% accuracy

(Thakre et al., 2022), gait analysis hit 95.6% (Susanty et al., 2023), a speech-based tool scored 95.27% (Ding et al., 2025), a text-based diary analysis achieved 85.0% (Shin et al., 2024), and acoustic features in Chinese students yielded 84.6% accuracy (Wei et al., 2025). These figures highlight AI's strong performance across diverse data types.

Sensitivity (correctly identifying individuals with depression) ranged from 69.0% (for a Bayesian Network prescreening tool) (Maekawa et al., 2024) to 96.53% (for emotional speech analysis) (Ding et al., 2025), indicating varied capabilities in case detection. Specificity (correctly ruling out those without depression) ranged from 68.0% (same Bayesian tool) (Maekawa et al., 2024) to 95.7% (gait analysis) (Susanty et al., 2023), reflecting differing abilities in avoiding false positives. Other key metrics included the Area Under the Curve (AUC), which measures overall diagnostic performance, ranging from 0.704 (for vocal features in older adults) (Georgescu et al., 2024) to 0.970 (for gait analysis) (Susanty et al., 2023). F1-scores, balancing precision and recall, spanned 0.5985 (for gait and rating scale fusion) (Liu et al., 2025) to 0.998 (for text and speech) (Hong et al., 2025). One study even reported a Spearman's r of 0.742 for ChatGPT-4 compared to PHQ-9, demonstrating strong alignment between AI and traditional tools (Liu et al., 2025). Every study compared their AI tools to trusted standards, such as PHQ-8, PHQ-9, BDI-II, GDS-15, GAD-7, or clinical diagnoses using DSM-5 criteria, MINI, or SCID (e.g., Ding et al., 2025; Liu et al., 2024; Mazur et al., 2025). Many found their tools performed as well as or better than these standards, suggesting AI's potential for faster and more accurate future screening. Table 2 summarizes these metrics for each study.

Table 2: Performance Metrics

Study (Author, Year)	Type of AI Tool Used	Overall Correct (%)	Found Depression Correctly (%)	Correctly Ruled Out No Depression (%)	Overall Test Performance Score (AUC)
Johnson et al., 2025	GPT (Large Language Model)	Not reported (Correlational)	Not reported	Not reported	Not reported
Bastos et al., 2025	Machine Learning (E-SVR)	Not reported (Correlational)	Not reported	Not reported	Not reported
Wei et al., 2025	Machine Learning (AdaBoost)	84.6	84.2	85	Not reported

Liu et al., 2024	LGMF-GNN (a type of Deep Learning)	78.75	Not reported	Not reported	80.64
Marieke M. van Buchem et al., 2024	NLP classifiers (BERT)	Not reported	Not reported	Not reported	0.88 (BERT)
Maekawa et al., 2024	Bayesian Networks	Not reported	69-76	68-77	0.77-0.90
Byeon, 2023	Machine Learning (Random Forest)	74.2	Not reported	Not reported	0.811 (training)
Hong et al., 2025	TIMEX-D (Deep Learning)	99.17-99.81	Not reported	Not reported	Not reported
Huang et al., 2024	Not named (wav2vec 2.0 - a voice AI)	94.81-96.49	Not reported	Not reported	Not reported
Ding et al., 2025	PM-SED (a Deep Learning tool)	95.27	96.53	94.02	Not reported
Liu et al., 2025	ChatGPT-4 (Large Language Model)	Not reported	Not reported	Not reported	0.852 (PHQ-9)
Mazur et al., 2025	Kintsugi Voice (v1)	Not reported	71.3	73.5	Not reported
Lee et al., 2024	Deep Learning (ResNet)	78	Not reported	Not reported	Not reported
Jia et al., 2025	MHA-GCN_ViT (a type of Deep Learning)	89.03	89.04	90.16	Not reported
Liu et al., 2025	Machine Learning/Deep Learning (Text-guided)	Not reported	Not reported	Not reported	Not reported
Thakre et al., 2022	Deep Learning	96.88	96.3	95.33	Not reported
Liu et al., 2023	PRA-Net (a Deep Learning tool)	Not reported	Not reported	Not reported	Not reported
Vu et al., 2025	Machine Learning (LightGBM)	90	87	91	Not reported
Susanty et al., 2023	Machine Learning (XGBoost)	95.6	95.5	95.7	0.97
Georgescu et al., 2024	Machine Learning (LR, XGBoost)	69.7	Not reported	Not reported	0.704

Implementation Characteristics and User Experience

While not all studies reported on real-world implementation or user experience, available insights are valuable. Most tools required common devices like smartphones, computers, or wearables; speech-based tools needed quality microphones and internet (e.g., Abd-alrazaq et al., 2023; Lee et al., 2024). Some utilized readily available data, such as patient messages or national survey records, which could facilitate easier deployment in diverse settings (e.g., Byeon, 2023; Marieke M. van Buchem et al., 2024). This suggests that access to technology is crucial and understanding how to reach individuals without such devices presents an opportunity.

Some AI tools simplified screening for users. For instance, one tool used routine electronic health records (EHR), eliminating the need for extra surveys (Byeon, 2023), and another collected data passively, reducing user burden (Lee et al., 2024). Tools employing large language models, like AI chatbots, could automate interviews, potentially saving time for both patients and clinicians (Liu et al., 2025). These approaches represent a significant step toward less burdensome mental health checks. User feedback was generally positive where reported. One study found clinicians sometimes trusted AI predictions over self-reports when discrepancies arose (Johnson et al., 2025), indicating confidence in the technology. Another noted users found a conversational AI bot engaging and comfortable (Liu et al., 2025). However, many studies did not extensively explore user satisfaction or long-term usability. Most tools remained in research settings, requiring further validation or clinical trials before widespread adoption.

Quality Assessment (Risk of Bias) Findings

I assessed the quality and potential risk of bias for the 21 studies using the Joanna Briggs Institute (JBI) Critical Appraisal Tools for observational and diagnostic studies (Joanna Briggs Institute, 2017). Overall, studies demonstrated a low risk of bias, bolstering confidence in their results, though some areas for improvement were noted.

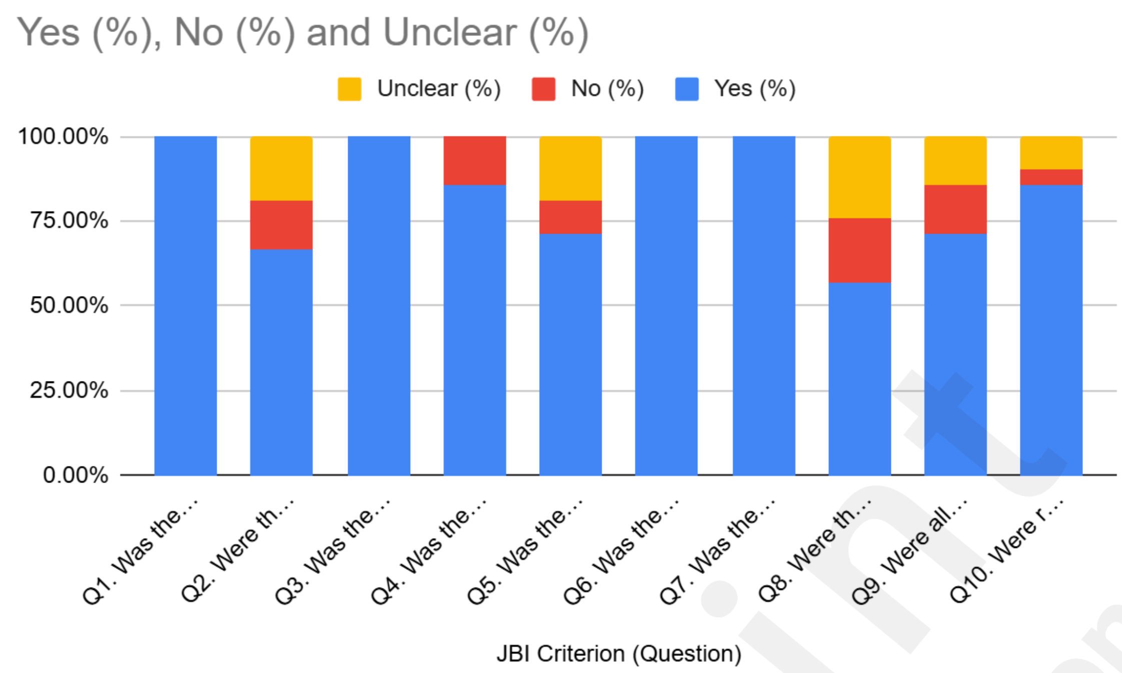
Studies generally excelled in clearly explaining goals, describing AI tools and their use, and relying on trusted standards like PHQ-9 or clinical diagnoses for comparison. They also appropriately timed AI and standard tests and kept results separate to avoid bias. Most provided robust details about participants, aiding in understanding the tested populations.

However, common challenges included insufficient explanation of original participant selection in studies using existing datasets, potentially introducing biases (e.g., convenience sampling). Some did not clarify if all participant data were included in analyses or if missing data impacted findings. A few omitted specific details like exact age ranges or the precise application of standard tools. While these gaps do not invalidate studies, they are important considerations for assessing result reliability. Figure 2 JBI Risk of Bias Assessment outlines these points.

Figure 2 JBI Risk of Bias Assessment:

JBI criterion questions:

1. Was there a clearly defined study question?
2. Were the inclusion and exclusion criteria for participants clearly described?
3. Was the index test (the AI tool) clearly described, and its application standardized?
4. Was the reference standard (the gold standard) clearly described, and its application standardized?
5. Was there an appropriate time interval between the index test and reference standard?
6. Was the reference standard interpreted independently of the index test?
7. Was the index test interpreted independently of the reference standard?
8. Were the participants recruited in a way that avoided bias?
9. Were all participants who received both the index test and reference standard included in the analysis?
10. Were relevant demographic and clinical characteristics of the participants described?



Study Limitations and Recommendations

The included studies highlighted several common limitations that could affect the generalizability and practical application of their findings. A recurring issue was the focus on specific cohorts, such as college students, older adults, or patients with particular conditions like cancer or diabetes. This limited applicability to broader populations, as a tool effective for young students, for example, might not translate well to older adults with different symptom presentations. Most studies also used cross-sectional designs, capturing data at a single point in time, which limited insights into depression's progression or the AI's ability to track it longitudinally. Furthermore, some studies relied on limited datasets, while others used only one data type (e.g., speech or gait), potentially missing other critical diagnostic cues. Studies that leveraged social media or patient messages sometimes lacked validation through a clinical diagnosis, raising concerns about their trustworthiness. Finally, many of the tools had not been tested in real clinical environments, leaving their practical efficacy and integration into routine healthcare unknown.

Based on these limitations, the studies collectively offered numerous valuable recommendations for future research. A key suggestion was to conduct longitudinal studies to track the evolution of depression and the AI's predictive capabilities over time. Another was to use larger, more diverse datasets to enhance tool reliability and ensure fairness across different ages, backgrounds, and health conditions. Integrating multimodal data—such as speech, gait, text, and brain scans—was also recommended to improve detection accuracy. For real-world impact, researchers suggested conducting trials in actual clinical settings to assess practical feasibility and patient-doctor acceptance. The studies also emphasized the need to improve AI model interpretability to build clinician trust and to prioritize ethical considerations like data privacy, security, and fairness. Finally, a clear recommendation was to evaluate different AI models to identify optimal solutions for screening.

Discussion and Conclusion

Discussion

My systematic review explored the effectiveness of AI-based tools in identifying and diagnosing depression in adults, delving into their strengths, features, and challenges. By synthesizing findings from 21 original studies published between January 1, 2022, and June 30, 2025, I found evidence that AI has significant potential to address limitations of traditional depression screening methods, offering substantial improvements in mental health care.

Effectiveness and Modalities of AI Tools

AI tools consistently demonstrated high effectiveness, often performing comparably to or better than traditional methods. Many achieved accuracies in the 80s and 90s, with one multimodal tool combining text and speech analysis reaching 99.81% accuracy (Hong et al., 2025). Sensitivity (correctly identifying individuals with depression) and specificity (correctly ruling out those without) were also strong, with some tools achieving sensitivity as high as 96.53% for emotional speech analysis (Ding et al., 2025) and specificity up to 95.7% for gait-based tools (Susanty et al., 2023). These figures suggest AI's reliability in identifying depression across various contexts.

The versatility of these tools is a key strength, as they leverage diverse data inputs to detect depression signals:

- **Speech:** Tools like Kintsugi Voice or wav2vec 2.0 analyze vocal cues from free-form or emotional speech (e.g., Ding et al., 2025; Huang et al., 2024; Mazur et al., 2025; Wei et al., 2025).
- **Text:** LLMs such as GPT-3.5 and ChatGPT-4 analyze diary entries, social media posts, or patient messages, demonstrating performance comparable to standard questionnaires like the PHQ-9 (e.g., Johnson et al., 2025; Liu et al., 2025; Marieke M. van Buchem et al., 2024; Shin et al., 2024).
- **Physical and Behavioral Signs:** Studies examined gait from wearables (e.g., Liu et al., 2025; Susanty et al., 2023), brain activity (EEG/fMRI) (Jia et al., 2025; Liu et al., 2024) , sleep patterns via polysomnography (Thakre et al., 2022), or facial expressions (Liu et al., 2023).
- **Health Records:** AI models were trained using existing electronic health records (EHR) or national survey data (e.g., Byeon, 2023; Maekawa et al., 2024; Vu et al., 2025).

This multimodal approach allows for less invasive and more objective screening by extracting insights from everyday behaviors or readily available data. For example, a gait-based tool could passively monitor mood via a smartwatch, offering a significant advancement over dense self-report surveys, which can be tiring, uncomfortable or hard for someone to access.

A crucial observation from Johnson et al. (2025) highlights a nuanced limitation for LLMs in depression screening. They found that while people might generate equally negative causal explanations for negative events, their underlying beliefs about the

changeability of those causes differ. It is this perceived changeability, not just the overall negativity, that predicts depressive symptoms. GPT, in their study, could not discern this changeability as effectively as the traditional paper-and-pencil Cognitive Style Questionnaire (CSQ). This suggests that while LLMs excel at processing natural language for qualitative insights, their current ability to interpret deeper cognitive constructs like "perceived changeability" for quantitative risk prediction remains a challenge. This underscores that human thought complexity, particularly subtle distinctions in meaning, may not always be accurately captured by current language models alone, even with extensive training on text.

Implementation of Characteristics and Public Health Implications

The potential for AI to transform mental health care is substantial, particularly in enhancing screening accessibility and reach. Smartphone or wearable-based tools could enable remote mental health checks, benefiting individuals in rural areas or those with limited access to mental health professionals (e.g., Abd-alrazaq et al., 2023; Lee et al., 2024). For instance, an EHR-based tool, like a nomogram, could seamlessly integrate into a doctor's routine without requiring additional effort (Byeon, 2023), while a gait-based tool might passively track an individual's mood (Susanty et al., 2023). These methods could facilitate earlier depression detection, leading to timely support and intervention before conditions worsen.

Studies also indicated that AI could alleviate burdens on both patients and clinicians. Tools that automate interviews or utilize existing data reduce the time and effort typically required by traditional surveys (e.g., Byeon, 2023; Liu et al., 2025). Where user feedback was reported, individuals generally responded positively to these tools—one study found users enjoyed a conversational AI bot engaging and comfortable (Liu et al., 2025). One study found clinicians sometimes trusted AI predictions over self-reports when discrepancies arose, indicating confidence in the technology (Johnson et al., 2025). However, many studies did not extensively explore user satisfaction or long-term usability. Most tools remained in research settings, requiring further validation or clinical trials before widespread adoption.

From a public health perspective, these tools could make mental health screening more scalable and affordable. Imagine clinics or community centers using AI to screen large populations efficiently, freeing up counselors to focus on treatment rather than assessments. This is particularly beneficial in underserved communities, where access to mental health care is often limited. The prospect of AI expanding mental health support availability and reducing its perceived overwhelming nature is truly inspiring.

Limitations of the Current Evidence Base

While the results are promising, several critical challenges remain:

- **Limited Generalizability:** Many studies focused on specific populations (e.g., college students, cancer patients, older adults), which may limit broad applicability (e.g., Bastos et al., 2025; Byeon, 2023; Marieke M. van Buchem et al., 2024; Susanty et al., 2023; Wei et al., 2025). Unclear participant selection in some public datasets also introduces potential for hidden biases.
- **Study Design:** The prevalence of cross-sectional designs limits insights into depression changes over time or AI's ability to predict future issues.
- **Data Quality:** Reliance on self-reported questionnaires, though valid, can be influenced by transient factors. Small datasets or non-clinically validated data (e.g., social media posts) raise reliability concerns.
- **Real-World Use:** Most tools remain in research settings, lacking comprehensive validation in busy clinical environments.
- **AI Challenges:** Complex deep learning models can lack interpretability for clinicians, potentially hindering trust. As noted in Johnson et al. (2025), LLMs, despite language processing prowess, may struggle with numerical tasks or discerning subtle emotional cues related to deeper cognitive constructs like "changeability".

Limitations of This Review

My review has its own limitations. I focused on studies from January 1, 2022, to June 30, 2025, to capture the latest AI advancements, potentially missing earlier foundational work. While major databases (PubMed, CINAHL, ScienceDirect) were searched, others (e.g., Scopus or PsycINFO) were not included, possibly overlooking some relevant studies despite broad search terms. As a single reviewer, despite adhering to strict guidelines, some subjectivity in screening and data extraction is inherent. Joanna Briggs Institute (JBI) tools were used for quality assessment, and ratings reflect my judgment.

Future Research Directions

Identified challenges suggest clear paths for future work to enhance AI tools:

- **Long-Term Studies:** Conduct longitudinal studies to track depression trajectories, predict future risks, and assess treatment outcomes over time.
- **Diverse Datasets:** Utilize larger, more varied datasets to enhance tool reliability across different ages, ethnicities, cultures, and health conditions, ensuring fairness and broad applicability.
- **Combining Data Types:** Integrate multimodal data (e.g., speech, gait, text, brain scans, smartphone activity) for more comprehensive and accurate depression detection.
- **Real-World Clinical Testing:** Conduct trials in actual clinical settings to evaluate practical integration, patient/doctor acceptance, and feasibility in daily practice.
- **Improved AI Interpretability:** Enhance AI model transparency to foster clinician trust and understanding of AI decision-making processes, encouraging wider adoption.
- **Ethical Considerations:** Prioritize data privacy, security, and fairness across diverse groups as AI development progresses, ensuring these tools do not exacerbate health disparities.
- **Comparative AI Model Analysis:** Evaluate various LLMs and other AI models to identify optimal approaches for screening depression, delving into their specific strengths and weaknesses

Conclusion

This review highlights the promise of AI-based tools for early depression detection in adults. They leverage diverse data modalities—from speech and gait to written text and health records—to achieve impressive results, often matching or exceeding traditional methods like questionnaires or clinical interviews. These tools could make mental health screening easier, faster, and more accessible, particularly for individuals with limited access to care. However, challenges persist, including limited generalizability, the need for longitudinal study designs, and the imperative for real-world clinical validation. By following the roadmap laid out for future research, such as robust clinical testing and the use of diverse data, we can unlock AI’s full potential to transform mental health care, ultimately helping more people receive timely support and improving their lives.

GRADE: Post-Hoc Analysis

To figure out how confident I can be in the overall evidence for AI tools screening for depression in adults, I used something called GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) principles. This approach helps me judge my confidence by looking at five main areas: how likely studies are to be biased, if results are consistent, if the studies are a good match for my question, how precise the findings are, and if there's any publication bias.

For the main thing I looked at – how accurate AI tools are for spotting depression – I rated my confidence as Low.

- Risk of Bias: While most individual studies seemed well-done (they had a low risk of bias according to my JBI checks), I still thought about this when deciding my overall confidence.
- Inconsistency: There was some difference in AI tools and people studied. For instance, Johnson et al. (2025) found that an old-fashioned paper test did better than an AI model in predicting depressive episodes, showing current AI limitations in picking up on tricky human thinking points like 'changeability,' which are important for predicting depression.
- Indirectness: Another concern was that many studies looked at very specific groups of people (like just college students or certain patient types). This means their findings might not apply well to everyone else. Also, some results were based on test scores, not a doctor's actual diagnosis.
- Imprecision: Most studies had clear performance figures and enough participants.

- Publication Bias: I couldn't formally check for this.

Because of this low confidence, even though AI tools show promise, we need more standardized and varied research before AI can be used widely in clinical environments.

Potential Conflicts of Interest

The author declares no potential conflicts of interest with respect to the research, authorship, or publication of this paper.

Funding Sources

No specific funding was received for this systematic review.

References

Abd-alrazaq, A., AlSaad, R., Aziz, S., Ahmed, A., Denecke, K., Househ, M., Farooq, F., & Sheikh, J. (2023). Wearable Artificial Intelligence for Anxiety and Depression: Scoping Review. *JMIR Mental Health, 25*(1), e42672.

Bastos, A. F., Fernandes-Jr, O., Liberal, S. P., Pires, A. J. L., Lage, L. A., Grichtchouk, O., Cardoso, A. R., de Oliveira, L. A., Pereira, M. G., De Boni, R. B., Volchan, E., & Erthal, F. S. (2025). Academic-related stressors predict depressive symptoms in graduate students: A machine learning study. *Behavioural Brain Research, 478*, 115328.

Byeon, H. (2023). Developing a nomogram for predicting depression in diabetic patients after COVID-19 using machine learning. *Frontiers in Public Health, 11*, 1150818.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication, 71*, 10-20. https://doi.org/10.1016/j.specom.2015.03.004

Ding, Z., Chen, J., Zhong, B. L., Liu, C. L., Liu, Z. T. (2025). Emotional stimulated speech-based assisted early diagnosis of depressive disorders using personality-enhanced deep learning. *Journal of Affective Disorders, 376*, 177-188.

Georgescu, A. L., Cummins, N., Molimpakis, E., Giacomazzi, E., Rodrigues Marczyk, J., & Goria, S. (2024). Screening for Depression and Anxiety Using a Nonverbal Working Memory Task in a

Sample of Older Brazilians: Observational Study of Preliminary Artificial Intelligence Model Transferability. *JMIR Formative Research, 8*, e55856.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports, 21*(11), 116. https://doi.org/10.1007/s11920-019-1094-0

Hong, J., Lee, J., Choi, D., & Jung, J. (2025). Depression level prediction via textual and acoustic analysis. *Computers in Biology and Medicine, 190*, 110009.

Huang, X., Wang, F., Gao, Y., Liao, Y., Zhang, W., Zhang, L., & Xu, Z. (2024). Depression recognition using voice-based pre-training model. *Scientific Reports, 14*(1), 12734.

Jacobson, N. C., Weingarden, H., & Wilhelm, S. (2020). Using digital phenotyping to accurately detect depression severity. *Journal of Nervous and Mental Disease, 208*(10), 768-775. https://doi.org/10.1097/NMD.0000000000001212

Jia, X., Chen, J., Liu, K., Wang, Q., & He, J. (2025). Multimodal depression detection based on an attention graph convolution and transformer. *Mathematical Biosciences and Engineering, 22*(3), 652-676.

Joanna Briggs Institute. (2017). *Critical appraisal tools*. https://jbi.global/critical-appraisal-tools

Johnson, J. K., & Haeffel, G. J. (2025). A paper-and-pencil questionnaire outperforms GPT for measuring cognitive vulnerability to depression and predicting depressive symptoms. *Scientific Reports, 15*(1), 19782.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., ... & Zaslavsky, A. M. (2015). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry, 21*(10), 1366-1371. https://doi.org/10.1038/mp.2015.198

Lee, T. R., Kim, G. H., & Choi, M. T. (2024). Geriatric depression and anxiety screening via deep learning using activity tracking and sleep data. *International Journal of Geriatric Psychiatry, 39*(2), e6071.

Liu, F., Wang, P., Hu, J., Shen, S., Wang, H., Shi, C., Peng, Y., Zhou, A. (2025). A psychologically interpretable artificial intelligence framework for the screening of loneliness, depression, and anxiety. *Applied Psychology: Health and Well-Being, 17*(1), 184–202.

Liu, X., Ren, M., Hu, X., Li, Q., & Huang, Y. (2025). Multimodal depression recognition based on gait and rating scale. *Expert Systems with Applications, 278*, 127285.

Liu, Z., Yuan, X., Li, Y., Shangguan, Z., Zhou, L., & Hu, B. (2023). PRA-Net: Part-and-Relation Attention Network for depression recognition from facial expression. *Computers in Biology and Medicine, 157*, 106589.

Liu, S., Zhou, J., Zhu, X., Zhang, L., Wang, G., & Jin, C. (2024). An objective quantitative diagnosis of depression using a local-to-global multimodal fusion graph neural network. *Patterns, 5*(12), 101081.

Liu, J., Gu, J., Tong, M., Yue, Y., Qiu, Y., Zeng, L., Yu, Y., Yang, F., & Zhao, S. (2025). Evaluating the agreement between ChatGPT-4 and validated questionnaires in screening for anxiety and depression in college students: a cross-sectional study. *BMC Psychiatry, 25*(1), 359.

Maekawa, E., Grua, E. M., Nakamura, C. A., Scazufca, M., Araya, R., Peters, T., & van de Ven, P. (2024). Bayesian Networks for Prescreening in Depression: Algorithm Development and Validation. *JMIR Mental Health, 11*(1), e52045.

Marieke M. van Buchem, A. A. H. de Hond, C. Fanconi, V. Shah, M. Schuessler, I. M. J. Kant, E. W. Steyerberg, & T. Hernandez-Boussard. (2024). Applying natural language processing to patient messages to identify depression concerns in cancer patients. *Journal of the American Medical Informatics Association, 31*(10), 2255-2262.

Mazur, A., Costantino, H., Tom, P., Wilson, M. P., & Thompson, R. G. (2025). Evaluation of an AI-Based Voice Biomarker Tool to Detect Signals Consistent With Moderate to Severe Depression. *Annals of Family Medicine, 23*(1), 15368-15369.

Mitchell, A. J., Vaze, A., & Rao, S. (2011). Clinical diagnosis of depression in primary care: A meta-analysis. *The Lancet, 374*(9690), 609-619. https://doi.org/10.1016/S0140-6736(09)60879-5

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews, 5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pękala, B., Garwol, K., Czuma, J., Kosior, D., Zaręba, L., & Chyła, M. (2023). Early detection of the risk of depressive episodes using a proprietary diagnostic test by new epistemic similarity measures. *Applied Soft Computing, 148*, 110910.

Shin, D., Kim, H., Lee, S., Cho, Y., & Jung, W. (2024). Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening Instrument Validation Study. *Journal of Medical Internet Research, 26*(1), e54617.

Susanty, S., Sufriyana, H., Su, E. C. Y., & Chuang, Y. H. (2023). Questionnaire-free machine-learning method to predict depressive symptoms among community-dwelling older adults. *PLoS ONE, 18*(1), e0280330.

Thakre, T. P., Kulkarni, H., Adams, K. S., Mischel, R., Hayes, R., & Pandurangi, A. (2022). Polysomnographic identification of anxiety and depression using deep learning. *Journal of Psychiatric Research*, *150*, 54-63.

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction, 27*(5), 1-53. https://doi.org/10.1145/3398069

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., ... & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry, 20*(3), 318-335. https://doi.org/10.1002/wps.20883

Villarroel, M. A., & Terlizzi, E. P. (2022). *Symptoms of depression among adults: United States, 2020*. NCHS Data Brief, (419), 1-8. https://www.cdc.gov/nchs/products/databriefs/db419.htm

Vu, T., Dawadi, R., Yamamoto, M., Tay, J. T., Watanabe, N., Kuriya, Y., Oya, A., Tran, P. N. H., & Araki, M. (2025). Prediction of depressive disorder using machine learning approaches: findings from the NHANES. *BMC Medical Informatics and Decision Making, 25*(1), 83.

World Health Organization. (2020). *Depression*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/depression

Zhao, Y., Hong, M., Shang, Y., & Yang, Y. (2021). Gait analysis for early detection of depression through deep learning. *Frontiers in Psychiatry*, *12*, 672564. https://doi.org/10.3389/fpsyt.2021.672564