Journal of Society Counseling
Vol. 4 No. 2, Agustus 2026

Review of Artificial Intelligence Applications in Mental Health Diagnosis and Therapy

Juster Donal Sinaga
Sanata Dharma University, Indonesia

Corresponding Author: donalsinaga@usd.ac.id

ARTICLE INFO
Article history:
Received
June 30, 2026
Revised
July 21, 2026
Accepted
July 24, 2026

ABSTRACT

Mental illness represents a growing global health concern, with increasing prevalence and burden across populations. Timely diagnosis and effective treatment are essential to improving individual health outcomes. In this context, artificial intelligence (AI) offers significant potential through rapid and precise data analysis. This narrative literature review aims to examine the application of AI technologies, specifically natural language processing (NLP), machine learning (ML), and predictive modelling, in the diagnosis and therapy of mental health disorders among adult clinical populations. The review includes twelve peer-reviewed articles indexed in Scopus between 2014 and 2024, selected based on relevance, methodological rigor, and contribution to the field. Findings were organised into three thematic clusters: (1) AI-assisted diagnostic accuracy and predictive modelling (e.g., fMRI-based PTSD prediction, depression detection via multimodal neural networks); (2) AI-enhanced therapeutic delivery and user engagement (e.g., AI-assisted online social therapy, remotely supervised brain stimulation); and (3) ethical, privacy, and implementation challenges (e.g., data bias, lack of transparency, and population representativeness). These technologies show promise in reducing human error and enhancing mental health care delivery; however, persistent challenges include data privacy, ethical considerations, and the need for diverse, large-scale datasets. Future research should focus on developing standardised implementation protocols using frameworks such as PRISMA, ensuring population diversity, and addressing ethical safeguards. Collaboration among mental health professionals, AI technologists, and policymakers is essential to promote safe, effective, and equitable integration of AI in psychological diagnosis and therapy.

Keywords: *Artificial Intelligence, Diagnosis, Treatment, Mental Health, Narrative Literature Review*

How to cite: Sinaga, J. D. (2026). Review of Artificial Intelligence Applications in Mental Health Diagnosis and Therapy. *Journal of Society Counseling*. 4(2). DOI: https://doi.org/10.59388/josc.v4i2.957
Journal Homepage: https://journal.scidacplus.com/index.php/josc

Published by: ScidacPlus

INTRODUCTION

Mental health is a global public health concern that demands sustained attention from researchers, policymakers, and practitioners alike. The World Health Organization (WHO, 2020) estimates that approximately one in four individuals will experience a mental or neurological disorder during their lifetime, underscoring the substantial burden of mental illness on individuals, families, and healthcare systems. Disorders such as depression, anxiety, bipolar disorder, and schizophrenia not only impair psychological functioning but also diminish quality of life, reduce productivity, and contribute to significant social and economic costs worldwide (Ritchie & Roser, 2018). Given the chronic and often progressive nature of many mental disorders, timely, accurate, and accessible diagnosis is essential for improving treatment outcomes and preventing long-term disability.

Traditional diagnostic approaches for mental health disorders rely primarily on self-reported symptoms and subjective clinical judgement, which are inherently prone to bias (Kilbourne et al., 2018). These methods are not only susceptible to variability in clinician expertise and experience, which may produce inconsistent diagnoses (Williams, 2022), but also lack the capacity for early detection, continuous monitoring, and the integration of complex behavioural and clinical information, all of which are critical for effective intervention (Carbonell et al., 2020). In contrast, artificial intelligence (AI) has the potential to complement conventional diagnostic practice by analysing large, heterogeneous datasets, identifying subtle patterns that may not be readily recognised through traditional assessment, and supporting more consistent and timely clinical decision-making (Shatte et al., 2019). Rather than replacing clinicians, AI can enhance diagnostic accuracy, facilitate earlier identification of individuals at risk, and enable continuous monitoring of mental health status over time. Nevertheless, social stigma continues to obstruct help-seeking behaviour, particularly among marginalised populations (WHO, 2022), indicating that technological advances must be accompanied by efforts to improve accessibility, trust, and equity in mental health care.

In response to these limitations, artificial intelligence (AI) has emerged as a transformative tool in the field of mental health care. AI-enabled technologies, including machine learning (ML), natural language processing (NLP), chatbots and virtual agents, digital phenotyping, and ecological momentary interventions, have demonstrated promising applications in both the diagnosis and management of mental disorders. ML algorithms can analyse large-scale datasets from electronic health records, social media platforms, and wearable devices to identify early indicators of mental disorders with greater consistency than traditional methods (Shatte et al., 2019). NLP-powered systems can analyse clinical notes and social media texts for linguistic patterns indicative of psychological conditions (D'Alfonso et al., 2017). Chatbots and AI-assisted platforms offer scalable, low-cost delivery of evidence-based psychological support (Koutsouleris et al., 2018). Digital phenotyping leverages data from smartphone usage, sleep patterns, and social interactions to model an individual's mental state in real time (Torous et al., 2016), while ecological momentary interventions deliver personalised, context-sensitive support based on continuously collected user data (Mohr et al., 2017).

Despite these promising developments, significant gaps remain in the literature. First, many existing studies focus narrowly on specific psychological conditions, predominantly depression and anxiety, without addressing a broader spectrum of mental health disorders (Guntuku et al., 2017). Second, findings from existing studies tend to demonstrate limited generalisability due to the use of non-diverse populations and constrained ecosystems (Chancellor & De Choudhury, 2020). Third, unresolved ethical and privacy concerns, particularly regarding consent, data security, and algorithmic bias, continue to pose major barriers to the widespread clinical adoption of AI in mental health care (Vayena et al., 2018).

These gaps give rise to the following focused research question: How can artificial intelligence be most effectively applied to the diagnosis and management of psychological conditions in clinical adult populations, and what are the key ethical and methodological challenges that must be addressed to ensure safe and equitable implementation?

This review pursues three specific objectives: (1) to evaluate the current state of AI applications in mental health diagnosis and treatment, with a focus on adult clinical populations; (2) to synthesise and thematically organise findings from recent empirical studies to assess the efficacy and limitations of various AI methodologies; and (3) to identify and critically examine gaps in the existing literature, particularly concerning population diversity, disorder range, ethical safeguards, and privacy considerations. These objectives provide a structured basis for deriving evidence-based recommendations for future research and clinical practice.

AI in the Diagnosis and Prognosis of Mental Health Disorders

A growing body of literature demonstrates the capacity of machine learning frameworks to enhance the accuracy and objectivity of mental health diagnosis. For instance, machine learning algorithms applied to functional magnetic resonance imaging (fMRI) data have shown significant potential for identifying neural correlates of post-traumatic stress disorder (PTSD) and depression, enabling pattern recognition across multimodal datasets that would be difficult to detect via conventional clinical assessment (Saeidnia et al., 2024; Thieme et al., 2020). These approaches offer the dual advantage of reducing dependence on subjective clinician judgment and enabling more consistent diagnostic outputs.

The use of deep learning and multimodal neural networks in detecting depression represents another significant advancement. Thieme et al. (2020) demonstrated that integrated audio-visual data and speech content could be used to identify depressive symptoms with minimal human involvement, pointing to AI's potential as a scalable screening tool for clinical settings. The predictive capacity of AI models has also been extended to more severe presentations of mental illness. Banerjee et al. (2021) reported that a class-contrastive ML approach was able to predict mortality risk among patients diagnosed with schizophrenia, yielding an AUROC score of 0.80, which reflects adequate discriminative ability. Collectively, these studies affirm that AI technologies can meaningfully complement traditional diagnostic processes, though their reliability is contingent upon the quality and representativeness of training datasets.

AI in Therapeutic Delivery and User Engagement

Beyond diagnosis, AI has demonstrated considerable utility in the delivery of therapeutic content and the facilitation of user engagement. D'Alfonso et al. (2017) found that AI-assisted online social therapy for young people recovering from psychosis effectively enhanced user participation and the relevance of therapeutic content delivery, suggesting that AI can extend the reach of evidence-based interventions beyond traditional clinical settings. This finding is complemented by Paneva et al. (2022), who demonstrated that remotely supervised transcranial electrical stimulation (TES), supported by telemedicine technologies, reduced both the cost and accessibility barriers associated with major depression treatment, thereby improving the potential for community-based care.

In addition, AI-driven platforms integrating cognitive behavioural therapy (CBT) have been developed to personalise therapeutic content through character-based storylines and adaptive delivery formats (Weaver et al., 2023). ChatGPT has also demonstrated performance exceeding that of the general population on emotional awareness tasks, suggesting potential applications in psychoeducational and supportive contexts (Elyoseph et al., 2023). These findings collectively indicate that AI can enhance not only the clinical effectiveness of therapy but also its reach, accessibility, and scalability across diverse populations.

Ethical, Privacy, and Implementation Challenges

Notwithstanding its promise, the clinical application of AI in mental health is associated with several critical ethical and implementation challenges. A recurring concern involves the risk of algorithmic bias arising from non-representative training datasets, which may amplify existing disparities in mental health care delivery (Rahman, 2023). Richmond et al. (2024) further highlighted specific privacy risks associated with AI-enabled voice technologies such as Amazon Alexa, which collect and process sensitive user data without sufficiently transparent consent mechanisms. Additionally, many AI models lack temporal structure and interpretability, complicating their integration into clinical decision-making workflows (Banerjee et al., 2021).

The broader ethical landscape also encompasses concerns about the completeness and representativeness of electronic health record (EHR) data used in training AI systems, the adequacy of informed consent processes, and the potential for data misuse (Vayena et al., 2018). An umbrella review by Abd-Alrazaq et al. (2022) concluded that while AI demonstrates reasonable diagnostic performance across disorders such as Alzheimer's disease, schizophrenia, and bipolar disorder, accuracy levels vary substantially, reinforcing the need for standardised evaluation protocols. These challenges underscore the imperative for clear regulatory frameworks, ethical guidelines, and inclusive research designs that prioritise both efficacy and equity in the clinical deployment of AI.

METHOD

This study adopts a narrative literature review methodology to systematically examine the application of AI technologies in the diagnosis and treatment of mental health disorders. Narrative reviews are appropriate when the goal is to synthesise a broad and heterogeneous body of literature within a defined thematic focus, rather than to conduct a formal meta-analysis (Snyder, 2019). This approach was selected over a systematic review given the diversity of AI methodologies and mental health contexts represented in the literature; however, efforts were made to introduce transparency and rigour into the selection process by following structured reporting guidelines adapted from the PRISMA framework (Moher et al., 2009).

The literature search was conducted using the Scopus database, which was selected for its comprehensive coverage of peer-reviewed, internationally indexed journals in psychology, medicine, and computer science (Falagas et al., 2008). The search encompassed articles published between 2014 and 2024 to capture both foundational and recent developments in the field. The following search terms were applied in combination using Boolean operators: "artificial intelligence", "mental health", "diagnosis", "treatment", "machine learning", "natural language processing", and "deep learning".

Inclusion criteria were defined as follows: (1) articles published in peer-reviewed journals indexed in Scopus between 2014 and 2024; (2) studies addressing the application of AI or ML to the diagnosis, prognosis, or treatment of mental health disorders; (3) articles written in English; and (4) studies utilising empirical, experimental, or systematic methodologies. Exclusion criteria included: (1) grey literature, conference proceedings, and non-peer-reviewed sources; (2) studies focusing exclusively on physical health conditions without any mental health component; (3) theoretical articles without empirical data; and (4) studies with methodological quality scores below the threshold established during screening.

The initial database search yielded 152 articles meeting preliminary keyword criteria. These articles were subjected to a two-stage screening process: title and abstract screening (Stage 1), followed by full-text assessment for eligibility (Stage 2). Following Stage 1, 110 articles were excluded based on irrelevance to the research question. The remaining 42 articles were assessed in full, and 12 were ultimately selected for inclusion based on their methodological rigour, relevance to the defined research question, and contribution to the identified thematic areas. Figure 1 presents the selection process in accordance with the simplified PRISMA-adapted flow.

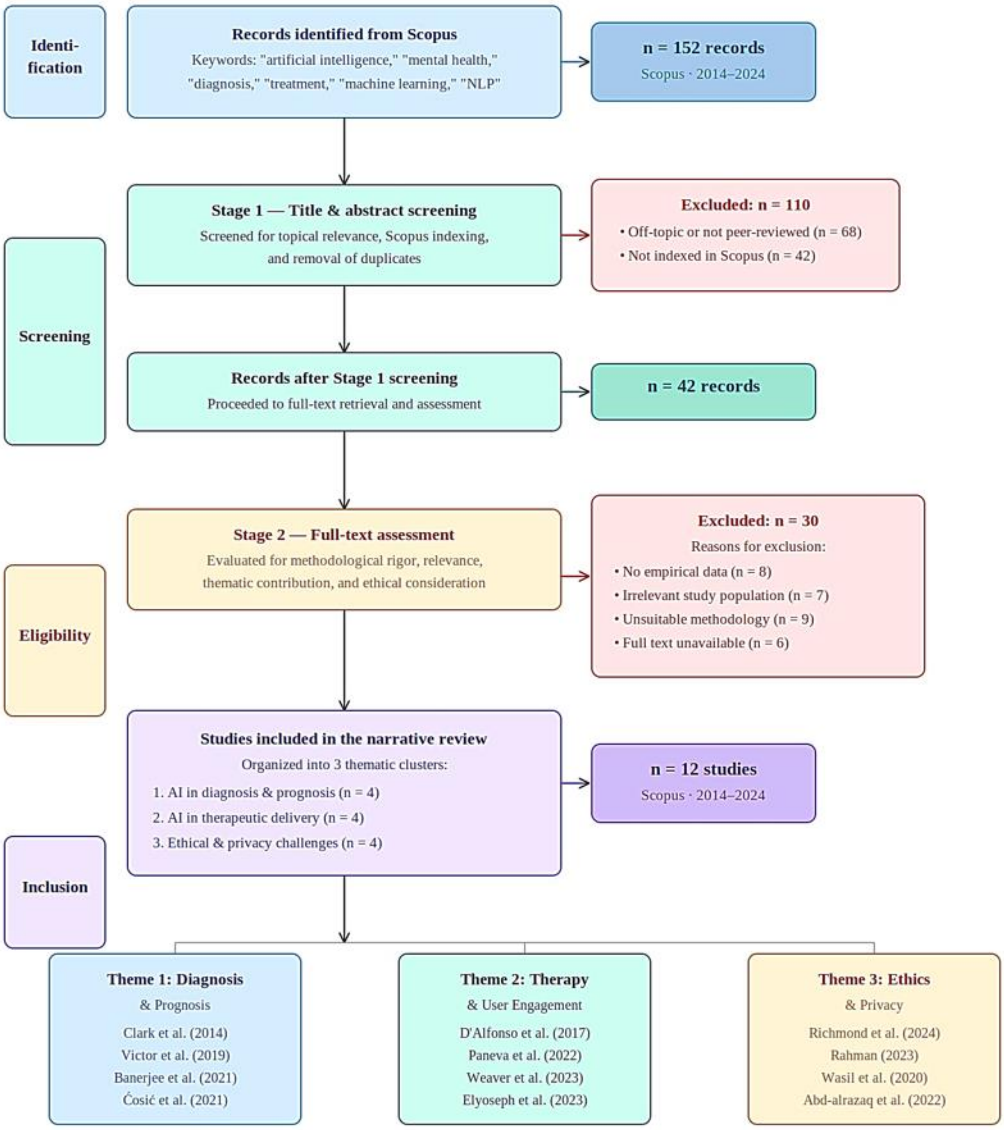

Figure 1. PRISMA flow diagram illustrating the literature search and article selection process.

Data were extracted from each included study using a standardised extraction template capturing the following variables: study title, journal name, year of publication, population studied, AI technologies employed, mental health condition(s) addressed, key advantages, and key limitations. Qualitative content analysis was then applied to identify recurring themes across studies, which were subsequently organised into three thematic clusters: (1) AI in diagnosis and prognosis; (2) AI in therapeutic delivery; and (3) ethical, privacy, and implementation challenges.

This structured approach facilitated systematic comparison across studies and ensured that the synthesis was grounded in evidence rather than impressionistic commentary.

RESULTS AND DISCUSSION

The findings of the study are presented thematically, drawing upon the 12 included studies summarised in Table 1. Results are organised according to three thematic clusters: (1) advanced AI applications in diagnosis, prognosis, and treatment; (2) gaps and limitations in the current literature on AI and psychological health; and (3) ethical and privacy concerns associated with AI use in mental health care.

Theme 1: Advanced AI Applications in the Diagnosis, Prognosis, and Treatment of Psychological Disorders

The first thematic cluster highlights the diagnostic and prognostic potential of AI in mental health care. Clark et al. (2014) utilised fMRI data to predict intrusive memories associated with PTSD, demonstrating that AI could identify specific brain networks involved in memory formation with approximately 68% accuracy across participants and up to 97% accuracy at the individual level. While this study was constrained by a small sample size and the use of a trauma film paradigm, which may not fully represent real-world traumatic experiences, its findings establish a significant proof of concept for AI-assisted neuroimaging in clinical psychiatry.

Victor et al. (2019) advanced this line of inquiry by demonstrating that deep learning and multimodal neural networks could detect depression with minimal human intervention, using integrated audio-visual data and speech content. Despite limitations related to data collection outside clinical environments and reliance on self-reported PHQ-9 measures, this study affirms AI's potential as a low-cost, scalable screening tool. D'Alfonso et al. (2017) further demonstrated that AI could significantly enhance user engagement and the delivery of therapeutic content within an online social therapy platform designed for youth recovering from early-stage psychosis. Although usage scale remained limited and automated linguistic analysis presented challenges in content relevance, this study highlights AI's capacity to bridge gaps in mental health service accessibility.

Beyond differences in reported performance, the reviewed studies suggest that AI diagnostic accuracy is highly dependent on methodological and contextual factors rather than on the AI algorithms alone. Variations in data quality, sample size, population characteristics, feature selection, and validation strategies substantially influence model performance. For example, Clark et al. (2014) achieved high individual-level accuracy using controlled fMRI data, whereas Victor et al. (2019) relied on multimodal audio-visual data collected in less controlled environments, introducing greater variability and reducing ecological consistency. Likewise, studies employing highly homogeneous datasets or narrowly defined clinical populations often report higher predictive accuracy than those involving more heterogeneous real-world populations, where symptom presentation and contextual factors are considerably more complex. These findings indicate that differences in AI performance primarily reflect differences in data representativeness, methodological design, and clinical context rather than the superiority of a particular AI technique. Consequently, future research should place greater emphasis on robust external validation, larger and more diverse datasets, and transparent reporting standards to improve the generalisability and clinical reliability of AI-assisted mental health diagnosis.

Collectively, these findings indicate that, while AI holds substantial promise across diagnostic, prognostic, and therapeutic dimensions, realising this potential in clinical practice will require addressing persistent limitations related to sample size, ecological validity, and the quality of training data.

Theme 2: Gaps in Current Studies on AI and Psychological Health

The second thematic cluster identifies significant gaps in the existing literature on AI and mental health, spanning three dimensions: disorder scope, population representativeness, and methodological rigour. With respect to disorder scope, several included studies focus narrowly on well-researched conditions such as depression, anxiety, and PTSD, while neglecting a broader range of mental health disorders (Guntuku et al., 2017). Wasil et al. (2020) observed that not all widely used mental health applications incorporate evidence-based content, and the effectiveness of AI-driven platforms may differ substantially depending on the clinical profile of the user. This concern points to the need for condition-specific evaluation frameworks rather than generic performance metrics.

With respect to population diversity, Richmond et al. (2024) observed that access to and effective use of AI technologies, including voice-based applications such as Amazon Alexa, is conditioned by users' technological literacy, language proficiency, and cultural context. These factors suggest that AI tools developed and validated within WEIRD (Western, Educated, Industrialised, Rich, Democratic) populations may not generalise effectively to more diverse global contexts (Chancellor & De Choudhury, 2020). A more deliberate emphasis on inclusive sampling procedures is therefore needed in future AI and mental health research.

With respect to methodological rigour, the included studies reveal variability in the transparency of AI model construction, validation procedures, and reporting standards. Banerjee et al. (2021) noted that predictive models frequently lack temporal structure and are sensitive to imbalances in training data, which compromises both reliability and interpretability. These findings highlight the need for standardised evaluation criteria and reporting frameworks, such as PRISMA, to enable meaningful comparison across studies.

Theme 3: Ethical and Privacy Concerns Associated with AI Utilisation in Mental Health

The third thematic cluster addresses the ethical and privacy implications of AI integration in mental health care. This constitutes one of the most pressing challenges in the field, given the sensitivity of mental health data and the potential consequences of misuse or misclassification.

Richmond et al. (2024) documented privacy concerns specific to AI-enabled voice technologies, highlighting that such systems may collect sensitive personal data from users, including minors, without adequately transparent consent mechanisms. This concern is compounded by findings from Rahman (2023), who demonstrated that AI systems trained on non-representative datasets risk amplifying existing social biases and perpetuating systemic discrimination against marginalised groups. The intersection of algorithmic bias and mental health care delivery is particularly consequential, as diagnostic and treatment errors in this domain carry significant personal and clinical ramifications.

Furthermore, Banerjee et al. (2021) identified transparency and interpretability as critical unresolved challenges in AI-driven mental health tools. When AI prediction models function as "black boxes", clinicians are unable to critically evaluate the basis of algorithmic recommendations, which erodes professional accountability and informed decision-making. In response to these concerns, the literature calls for the development and enforcement of clear, jurisdiction-specific guidelines governing the ethical use of AI in mental health contexts, including standards for data governance, consent, bias auditing, and algorithmic accountability

(Vayena et al., 2018). Such guidelines are essential preconditions for the safe, equitable, and effective integration of AI in clinical mental health practice.

Table 1. Summary of the Literature Review on the Implementation of AI in Mental Health Diagnosis and Treatment

No	Title	Population	Technologies Used	Health Issues	Advantages	Limitations
1	First steps in using machine learning on fMRI data to predict intrusive memories of traumatic film footage	57 participants from two separate studies	Functional Magnetic Resonance Imaging (fMRI)	Post-Traumatic Stress Disorder	Predicted intrusive memories with 68% accuracy (97% for individual participants); identified brain networks involved in intrusive memory formation.	Limited sample size; the trauma film paradigm may not fully replicate real trauma.
2	A non-randomized feasibility study of a voice assistant for parents to support their children's mental health	55 community parents and 4 clinical parents	Amazon Alexa, Automatic Speech Recognition, NLU, NLP	Children's mental health	Easy-to-use; potential for interactive parenting skills practice; hands-free use.	Difficulty with voice commands, privacy concerns, and robotic voice quality.
3	Detecting Depression Using a Framework Combining Deep Multimodal Neural Networks With a Purpose-Built Automated Evaluation	671 participants	Deep learning, multimodal neural networks, audio-visual data, Google Cloud APIs	Depression	Depression detection with minimal human intervention: potential as a clinical screening tool.	Data collected outside clinical settings; relies on self-reported PHQ-9 scores; requires further clinical validation.
4	Reassessing Evidence-Based Content in Popular Smartphone Apps for Depression and Anxiety	Top 27 most popular mental health apps	Apple App Store, Google Play Store, Mobile Action analytics	Depression and anxiety	Used monthly active user data to customise analysis; highlighted differences between user-customised and non-	User data from only one month (July 2018); not all popular apps use evidence-based content.

No	Title	Population	Technologies Used	Health Issues	Advantages	Limitations
					customised analysis.	
5	Machine Learning and Natural Language Processing in Psychotherapy Research: Alliance as Example Use Case	386 clients, 40 therapists at a university counselling centre	Kaldi (speech recognition), Python (scikit-learn), Sent2vec, tf-idf	Therapeutic alliance (not disorder-specific)	Utilises large recorded therapy session data; applies NLP and ML to predict therapy process variables; potential to automate therapy quality assessment.	Small sample size for ML standards; transcription and diarisation errors; retrospective alliance assessment may reduce accuracy.
6	Entertain Me Well: An Entertaining, Tailorable, Online Platform Delivering CBT for Depression	Case studies across various populations	Video-based online platform with character-based storylines	Depression	Character storylines increase engagement; low-cost content customisation; simple content delivery.	Currently only in English; limited to one-character storyline; untested without human support.
7	A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness	1,706 patients with schizophrenia	Machine learning, class-contrastive reasoning, autoencoder, random forest	Schizophrenia	Mortality prediction with AUROC 0.80; class-contrastive reasoning for interpretability; highlighted role of comorbidities.	No temporal structure in model; sensitivity to training data imbalance; limited generalisation to other severe mental illness types.
8	AI-Based Prediction and Prevention of Psychological and Behavioral Changes in Ex-COVID-19 Patients	Ex-COVID-19 patients	Machine learning, NLP, neurophysiology, facial/oculometric features	COVID-19-related mental health disorders (depression, anxiety, PTSD)	Integrates multimodal features; improves prediction accuracy; predicts and prevents mental health disorders.	Challenges in external validation; limited longitudinal data; needs further research on ethical and privacy concerns.
9	Artificial Intelligence-Assisted Online Social	Youth recovering from early psychosis	AI, NLP, chatbot, sentiment analysis	Psychosis, depression	Combines peer support with evidence-based interventions;	Limited current user scale; cannot fully replace human

No	Title	Population	Technologies Used	Health Issues	Advantages	Limitations
	Therapy for Youth Mental Health				enhances user engagement; integrates persuasive design for behaviour change.	moderators; challenges in linguistic analysis and automated content relevance.
10	Using Remotely Supervised At-Home TES for Enhancing Mental Resilience	Patients with major depressive disorder (MDD)	Transcranial Electrical Stimulation (TES), Transcranial Direct Current Stimulation, telemedicine	Depression, relapse prevention	Reduces accessibility barriers and costs; potential for personalisation and continuous monitoring.	Risk of overuse or underuse; variability in electrode placement; lack of long-term effects data.
11	ChatGPT outperforms humans in emotional awareness evaluations	General population	ChatGPT	Emotional awareness	ChatGPT outperforms the general population in emotional awareness; high accuracy in contextual emotion adjustment.	Interactions only in English with French-speaking normative data; produces illusory responses; limited to a specific AI model and period.
12	AI for ADHD: Opportunities and Challenges	Non-specific	Machine learning, deep learning, virtual reality	ADHD	Improves ADHD diagnosis and management via multi-source data analysis; provides 24/7 support; uses brain biomarkers for monitoring.	Risk of reinforcing bias if training data is not representative; privacy concerns; lack of transparency in AI prediction.

The present review synthesises evidence from 12 empirical studies to address the central question of how AI can be most effectively applied to the diagnosis and management of psychological conditions, and what challenges must be resolved to ensure safe and equitable implementation. The findings confirm that AI holds substantial potential across multiple dimensions of mental health care, while also revealing persistent limitations that must be addressed before widespread clinical adoption is feasible.

With respect to diagnostic accuracy, the evidence reviewed supports the utility of machine learning frameworks, particularly fMRI-based models and multimodal neural networks, in enhancing the objectivity and precision of mental health diagnosis. Clark et al. (2014) and Victor et al. (2019) both demonstrated meaningful diagnostic performance, though each study was constrained by sample size limitations and ecological validity concerns. These constraints are well documented in the broader methodological literature: small samples reduce statistical power and limit generalisability to the target population, thereby compromising the external validity of findings (Cao et al., 2024). Future studies should therefore prioritise larger, more clinically representative samples and employ cross-validation designs to enhance the robustness of their findings.

With respect to therapeutic delivery, the AI-assisted platforms reviewed, including online social therapy systems, CBT delivery platforms, and remotely supervised brain stimulation technologies, demonstrate that AI can meaningfully extend the reach and effectiveness of evidence-based interventions beyond the constraints of traditional clinical settings (D'Alfonso et al., 2017; Paneva et al., 2022; Weaver et al., 2023). However, Balcombe (2023) appropriately cautioned that AI systems frequently struggle to comprehend the nuances of natural language and therapeutic context, which may result in inappropriate or irrelevant content being delivered to vulnerable users. Saeidnia et al. (2024) further emphasised that inclusive, representative training datasets are a prerequisite for AI systems that can respond meaningfully and equitably to the diverse needs of clinical populations.

With respect to ethical and population-related concerns, the review highlights a persistent gap in the literature regarding study population diversity. Many included studies were conducted with predominantly Western, English-speaking, or otherwise homogeneous samples, which raises fundamental questions about the generalisability of findings to non-WEIRD populations. Richmond et al. (2024) and Wasil et al. (2020) both emphasise the importance of culturally and linguistically inclusive designs in AI research. Banerjee et al. (2021) and Rahman (2023) further argue that robust bias auditing and privacy safeguards must be embedded within the design and deployment of AI systems, rather than treated as post-hoc considerations. Without such safeguards, AI tools risk exacerbating existing inequities in mental health care access and quality.

These converging findings underscore the need for interdisciplinary collaboration among mental health practitioners, AI developers, ethicists, regulators, and the communities most affected by mental health disparities. The integration of AI into clinical mental health practice must be guided by evidence-based standards, structured implementation protocols, and transparent accountability mechanisms to ensure that technology serves the therapeutic goals of accuracy, equity, and patient dignity.

CONCLUSION

This narrative literature review has synthesised evidence from 12 empirical studies to evaluate the application of artificial intelligence in the diagnosis and treatment of mental health disorders. The review confirms that AI technologies, including machine learning, deep learning, NLP, and AI-assisted therapeutic platforms, demonstrate meaningful potential to improve diagnostic accuracy, expand therapeutic accessibility, and personalise mental health care delivery. Studies such as Clark et al. (2014) and Victor et al. (2019) have established proof of concept for AI-driven diagnosis of PTSD and depression, while D'Alfonso et al. (2017) and Paneva et al. (2022) illustrate the capacity of AI to enhance therapeutic engagement and reduce barriers to care.

At the same time, the review identifies several significant limitations that constrain the clinical readiness of current AI applications. These include small and non-diverse study populations, limited ecological validity, the absence of standardised evaluation frameworks, and unresolved ethical concerns pertaining to algorithmic bias, data privacy, and model transparency. The literature consistently calls for the adoption of structured review frameworks (PRISMA), the inclusion of diverse and representative populations in AI research, and the development of clear regulatory guidelines governing the ethical use of AI in clinical mental health contexts.

Future research should prioritise large-scale, multi-site studies with clinically and demographically diverse samples; adopt transparent, reproducible methodologies; and pursue meaningful collaboration among mental health professionals, AI developers, and policymakers. Only through such coordinated and ethically grounded efforts can AI be integrated into mental health care in a manner that is safe, equitable, and clinically effective.

REFERENCES

Abd-Alrazaq, A., Alhuwail, D., Schneider, J., Toro, C. T., Ahmed, A., Alzubaidi, M., Alajlani, M., & Househ, M. (2022). The performance of artificial intelligence-driven technologies in diagnosing mental disorders: An umbrella review. *npj Digital Medicine*, 5, Article 87. https://doi.org/10.1038/s41746-022-00631-8

Banerjee, S., Lio, P., Jones, P. B., & Cardinal, R. N. (2021). A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness. NPJ Schizophrenia, 7, Article 61. https://doi.org/10.1038/s41537-021-00181-0

Balcombe, L. (2023). AI chatbots in digital mental health. *Informatics, 10*(4), Article 82. https://doi.org/10.3390/informatics10040082

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cao, Y., Chen, R. C., & Katz, A. J. (2024). Why is a small sample size not enough? *The Oncologist, 29*(9), 761–763. https://doi.org/10.1093/oncolo/oyae105

Carbonell, Á., Navarro-Pérez, J. J., & Mestre, M. V. (2020). Challenges and barriers in mental healthcare systems and their impact on the family: A systematic integrative review. *Health & Social Care in the Community, 28*(5), 1366–1379. https://doi.org/10.1111/hsc.12968

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine*, 3, Article 43. https://doi.org/10.1038/s41746-020-0233-7

Clark, I. A., Niehaus, K. E., Duff, E. P., Di Simplicio, M. C., Clifford, G. D., Smith, S. M., Mackay, C. E., Woolrich, M. W., & Holmes, E. A. (2014). First steps in using machine learning on fMRI data to predict intrusive memories of traumatic film footage. *Behaviour Research and Therapy*, 62, 37–46. https://doi.org/10.1016/j.brat.2014.06.006

Ćosić, K., Popović, S., Šarlija, M., Kesedžić, I., Gambiraža, M., Dropuljić, B., Mijić, I., Henigsberg, N., & Jovanovic, T. (2021). AI-based prediction and prevention of psychological and behavioral changes in ex-COVID-19 patients. *Frontiers in Psychology*, 12, Article 636228. https://doi.org/10.3389/fpsyg.2021.636228

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.)*. SAGE Publications.

D'Alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., Gleeson, J., & Alvarez-Jimenez, M. (2017). Artificial intelligence-assisted online social therapy for youth mental health. *Frontiers in Psychology*, 8, Article 796.

https://doi.org/10.3389/fpsyg.2017.00796

Dalgleish, T., Black, M., Johnston, D., & Bevan, A. (2020). Transdiagnostic approaches to mental health problems: Current status and future directions. *Journal of Consulting and Clinical Psychology, 88*(3), 179–195. https://doi.org/10.1037/ccp0000482

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14, Article 1199058. https://doi.org/10.3389/fpsyg.2023.1199058

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Fusar-Poli, P., Correll, C. U., Arango, C., Kelleher, I., Treasure, J., & Kessing, L. V. (2021). Preventive psychiatry: A blueprint for improving the mental health of young people. *World Psychiatry, 20*(2), 200–221. https://doi.org/10.1002/wps.20869

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., Villatte, J. L., Georgiou, P. G., Van Epps, J., Imel, Z. E., Narayanan, S. S., & Atkins, D. C. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology*, 67(4), 438–448. https://doi.org/10.1037/cou0000382

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

Kilbourne, A. M., Beck, K., Spaeth-Rublee, B., Ramanuj, P., O'Brien, R. W., Tomoyasu, N., & Pincus, H. A. (2018). Measuring and improving the quality of mental health care: A global perspective. *World Psychiatry, 17*(1), 30–38. https://doi.org/10.1002/wps.20482

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Dwyer, D. B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., Schmidt, A., Gillam, J., Schultze-Lutter, F., Falkai, P., Reiser, M., Riecher-Rössler, A., Upthegrove, R., Hietala, J., Salokangas, R. K. R., Pantelis, C., … Schroeder, J. (2018). Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry*, 75(11), 1156–1172. https://doi.org/10.1001/jamapsychiatry.2018.2165

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *BMJ*, 339, Article b2700. https://doi.org/10.1136/bmj.b2700

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, 13, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. *The New England Journal of Medicine, 375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Paneva, J., Leunissen, I., Schuhmann, T., de Graaf, T. A., Jønsson, M. G., Onarheim, B., & Sack,

A. T. (2022). Using remotely supervised at-home transcranial electrical stimulation (TES) for enhancing mental resilience. *Frontiers in Human Neuroscience*, 16, Article 838187. https://doi.org/10.3389/fnhum.2022.838187

Rahman, M. M. (2023). AI for ADHD: Opportunities and challenges. *Journal of Attention Disorders, 27*(8), 797–799. https://doi.org/10.1177/10870547231160504

Richmond, S., Bell, M., Ngo, D., & Yap, M. B. H. (2024). A non-randomized feasibility study of a voice assistant for parents to support their children's mental health. *Frontiers in Psychology*, 15, Article 1390556. https://doi.org/10.3389/fpsyg.2024.1390556

Ritchie, H., & Roser, M. (2018). *Mental health. Our World in Data*. https://ourworldindata.org/mental-health

Saeidnia, H. R., Hashemi Fotami, S. G., Lund, B., & Ghiasi, N. (2024). Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact. *Social Sciences, 13*(7), Article 381. https://doi.org/10.3390/socsci13070381

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

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

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), Article 34. https://doi.org/10.1145/3398069

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. JMIR Mental Health, 3(2), e16. https://doi.org/10.2196/mental.5165

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine, 15*(11), Article e1002689. https://doi.org/10.1371/journal.pmed.1002689

Victor, E., Aghajan, Z. M., Sewart, A. R., & Christian, R. (2019). Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. *Psychological Assessment, 31*(8), 1019–1031. https://doi.org/10.1037/pas0000822

Wasil, A. R., Gillespie, S., Patel, R., Petre, A., Venturo-Conerly, K. E., Shingleton, R. M., Weisz, J. R., & DeRubeis, R. J. (2020). Reassessing evidence-based content in popular smartphone apps for depression and anxiety: Developing and applying user-adjusted analyses. *Journal of Consulting and Clinical Psychology, 88*(11), 983–993. https://doi.org/10.1037/ccp0000492

Weaver, A., Zhang, A., Xiang, X., Felsman, P., Fischer, D. J., & Himle, J. A. (2023). Entertain me well: An entertaining, tailorable, online platform delivering CBT for depression. *Cognitive and Behavioral Practice, 30*(1), 1–14. https://doi.org/10.1016/j.cbpra.2021.10.004

Williams, D. R. (2022). Stress and the mental health of populations of color: Advancing our understanding of race-related stressors. *Journal of Health and Social Behavior, 63*(1), 3–

19. https://doi.org/10.1177/00221465211073027
World Health Organization. (2020). Mental health: Strengthening our response. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response.
World Health Organization. (2022). Mental health and COVID-19: Early evidence of the pandemic's impact. https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1