

Evaluating AI Models for Mental Health Monitoring and Diagnosis: A Comprehensive Systematic Review across Text, Speech and Physiological Signals

Gaurav Sharma¹[0009-0008-6358-1259] and Gaurav Gupta²[0000-0002-5192-4428]

¹Shoolini University, Solan, Himachal Pradesh 173229, India

²Shoolini University, Solan, Himachal Pradesh 173229, India

Abstract. The scale of the global mental health crisis — affecting roughly 970 million people worldwide stands in sharp contrast to the tools we currently have for detecting and monitoring it. Questionnaires are subjective. Clinical interviews are resource-intensive. And in much of the world, there simply are not enough trained professionals to meet demand. Over the past decade, artificial intelligence has been proposed as a way to address some of these gaps, and a substantial body of research has accumulated around three main data sources: physiological signals, text, and speech. This paper reports findings from a PRISMA-guided systematic review in which we searched eight major databases including PubMed, IEEE Xplore, ACM Digital Library, and ScienceDirect for peer-reviewed studies published between 2015 and 2025. After two independent reviewers screened 120 records, 63 met our inclusion criteria. We grouped these into four modality categories: physiological signal studies (EEG, HRV, ECG, electrodermal activity), NLP-based text analysis, speech and acoustic analysis, and multimodal fusion approaches. EEG-based deep learning models reported the highest accuracy figures sometimes above 97%, though we argue these numbers are difficult to interpret without knowing the sample size and experimental conditions behind them, which in most cases were quite limited. Transformer models such as BERT and RoBERTa did well on depression and suicidality detection tasks, speech models showed useful diagnostic signals via acoustic and prosodic features, and multimodal systems outperformed single-source approaches across the board. What the review also surfaces, though, is a consistent set of problems: small and unrepresentative samples, almost no external validation, very little attention to fairness or interpretability, and reporting practices that make replication difficult. These gaps matter. They are what currently stands between technically interesting research and tools that could actually be used in clinical practice.

Keywords: Anxiety, depression, mental health, machine learning, deep learning, artificial intelligence.

(Article history: Selected papers from NIELIT's International Conference on Communication, Electronics and Digital Technologies (NICE-DT'26), 29-30 Jan, 2026, Gorakhpur, India)

1 Introduction

Mental health disorders represent one of the most pressing global health challenges, that affects over 970 million people across the globe [3, 19] and contributes significantly to disability-adjusted life years [19, 1]. Depression alone impacts more than 264 million individuals [19], while global prevalence of anxiety disorders has been reported ranging from 4% to 16% [28, 5]. The COVID-19 pandemic has further increased these numbers, with notable increases in depressive symptoms, anxiety, and distress across the world [19, 21].

Self-report questionnaires, clinical interviews, and periodic consultations are heavily used in traditional mental health assessment which are described as inherently subjective, potentially time-consuming, and prone to patient biases and interpretation [11, 12]. These limitations, combining the global shortage of mental health professionals as some sources noted that only 1% of the healthcare workforce provides mental health services across the world [18] and accessibility barriers (including long wait times and high costs) [42, 1, 30], create significant gaps in early detection and continuous monitoring of mental health conditions [11, 17, 20].

The widespread use of digital technologies has opened new ways to objectively and continuously keep an eye on mental health [11, 13]. Smartphones, social media sites, wearable tech, and biosensors all collect a lot of behavioural and physiological data that can be used as digital biomarkers of mental health [7, 12]. Artificial intelligence (AI) and machine learning (ML) techniques are essential to analyze these complex, multi-dimensional data streams [7, 21, 22]. AI is specifically noted for its capability to identify specific non-obvious patterns associated with mental health conditions that may be imperceptible to human observers [3, 28, 32].

Three primary modalities have emerged as particularly promising for AI-based mental health detection: text analysis of social media posts, clinical transcripts, and digital communications utilizing Natural Language Processing (NLP) to analyze written text, often from social media or online forums, to identify linguistic cues and sentiment indicative of mental disorders [33, 14]; speech analysis using acoustic and linguistic features from voice recordings leveraging AI to analyze acoustic, prosodic,

and linguistic features such as tone, pitch, pace, and volume to detect affective cues and early signs of mental illness [21, 28]; and physiological signal analysis from wearable sensors measuring heart rate variability (HRV), electroencephalography (EEG), and other bio signals employing wearable and biosensor devices to continuously collect objective data like HRV, Electroencephalography (EEG), and electrodermal activity (EDA/GSR) to monitor physiological reactions correlated with stress and anxiety [12, 21].

Recent progress in deep learning, especially transformer-based models such as BERT and RoBERTa, has demonstrated exceptional efficacy in natural language processing (NLP) tasks pertinent to mental health [14, 16]. These models are esteemed for their capacity to elucidate intricate linguistic patterns linked to conditions such as depression and suicidal ideation [27]. For example, RoBERTa has been shown to work better than traditional classifiers, finding depression with high accuracy on sites like Twitter [27]. Likewise, transformer-based language models are employed for the preliminary identification of depressive tendencies in textual data [19, 22]. For speech signals, advanced AI frameworks use deep learning and large language models (LLMs) to look at acoustic, prosodic, and linguistic features. This makes it easier to find emotional expressions [12]. LLM-based models show that they can do things like evaluate disorganized thought processes (like in schizophrenia) with the same level of accuracy as expert ratings [29]. Specialized deep learning models trained on speech data can also tell the difference between depressive and non-depressive speech with "impressive accuracy" [30]. Ensemble learning methods and convolutional neural networks (CNNs) have shown great accuracy in classifying mental health states from EEG and HRV data [12]. CNNs are great for working with high-dimensional, multi-dimensional time-series data, and in some cases, they can get accuracy rates of up to 92% [1]. Moreover, hybrid architectures that combine CNNs (for feature extraction) with recurrent networks (for sequence modeling), like CNN-BiLSTM, are often praised for greatly improving prediction accuracy compared to single model architectures [12, 1].

Even with these encouraging advancements, numerous essential inquiries persist concerning the clinical preparedness of AI mental health systems [1, 23]. Performance metrics reported in research settings may not translate to real-world effectiveness due to a reliance on small datasets, lack of diversity in training data, and the absence of proper external validation procedures [12, 22]. The Literature very often inadequately represents fairness, interpretability, and calibration—crucial for clinical decision-making and fostering trust causing concerns [12, 23]. It makes hard for doctors to use the system because they need to know how decisions are made in order to make a diagnosis or prediction; this problem is known as “black box problem”. [12, 27].

In this review we seek to deliver a thorough assessment of AI model efficacy across textual, auditory, and physiological modalities for mental health surveillance and diagnosis. To see what clinical translation can and can't do right now and what it needs to do we look at model architectures, performance metrics, validation methods, and reporting quality to figure out.

2 Literature Review

Research on the use of Artificial Intelligence (AI) and Machine Learning (ML) in mental health has expanded at a remarkable pace in last few years, with studies making use of resources from different types of data, methodological traditions, and technological tools. Across this literature, we find a shared effort to understand how digital systems can support monitoring, diagnosis, and therapeutic decision-making. Many researchworks also pay close attention to issues of accuracy, transparency, and ethical safeguards. They have become central considerations as AI is integrate more deeply into real clinical environments [7, 29, 13, 16, 22].

One of themotivations behind these studies is the big scale of the global mental-health burden. The analysis by Mijeong Kang highlights the immense economic burden of mental illness, which is anticipated to reach USD 16 trillion by 2030 [8, 6, 9, 5], and it emphasizes how this crisis has spurred the advancement of innovative sensing technologies. For example,for continuous physiological monitoring hardware and mental-health research are becoming increasingly interconnected through well designed machine-knitted wearable technology [5, 12]. The emerging materials used in these devices comply with established standards for electrodermal measurement, which indicates that technological novelty is being matched with scientific rigor.

Several studies reflect the move toward continuous and remote monitoring. As described by Francesco Monaco and colleagues in the Innova4Health project, mobile and wearable devices [7] are used to track symptoms of Major Depressive Disorder. Strong emphasis on ethical governance stands out here, this Innova4Health project study's protocol and consent procedures were reviewed and approved by the Campania Sud Review Board. Shaik and Nigar, who explore remote patient monitoring platformsusing wearable sensors, passive RFID, and wireless communication have shown similar systems [15]. Their findings have highlighted both the potential of these systems and the practical challenges related to privacy, interoperability, and data security.

The second big problem is AI-based diagnostic and predictive modelling. If we see Jiao's SA-BFRNN (Snow Ablation-driven Bi-directional Fine-tuned Recurrent Neural Network) which is a model intended for monitoring mental health via biosensor data (e.g., heart rate variability and skin conductance) and significantly surpasses conventional methods such as Logistic Regression (LR: 79.89%) and K-Nearest Neighbors (KNN: 76.27%), attaining a precision rate of 89.23% [10]. (The outcomes

of this model were optimal in both training and testing phases). Li and his co-authors have also looked into quantum-enhanced machine learning methods for finding psychiatric biomarkers in EEG data [11]. Their new and improved Prairie Dog Optimized Poly-Kernel Support Vector Machine (RPDO-PSVM) method is 96% accurate, 93.8% precise, 92.1% recall, and 94.4% F1-score, which is better than older methods. This shows usefulness of hybrid, specialized algorithms and how combining biosensor technology with strategic training could improve mental health.

As predictive modeling is becoming more common, researchers have also become more vocal about the need for interpretability. According to Brando and colleagues clinical AI systems, particularly those used for sensitive applications such as suicide-risk estimation cannot rely solely on opaque “black-box” models [12]. This view is supported by Olatinwo’s work which demonstrates how interpretability tools like LIME and Shapley values can be integrated to make deep learning models easier for clinicians to understand. Studies by Omiyefa and Singh being ethics oriented similarly call for transparency, fairness, and human oversight aligning mental-health AI with international regulatory expectations [14, 16].

Natural language processing (NLP) research adds another strand to the literature. Zubair and colleagues show how deep learning models—including LSTMs, RNNs [14, 16], and more advanced transformer architectures—can detect suicide-related sentiment in social media data with high recall values. Importantly, they adopt a transfer-learning approach, training models first on clinical notes before adapting them to Twitter, resulting in an F-measure of 0.83 for the positive (suicide-associated) class. Complementary studies by Allam, Baydili, and Cosic reaffirm the value of both transformer models and more traditional classifiers like Support Vector Machines, which continue to perform strongly in several datasets related to depression, anxiety, and PTSD [32, 29, 24].

A growing subset of the literature examines conversational agents and digital therapeutic platforms. Randomized trials reported by Rosenfeld, D’Alfonso, Jadhav, and K.P. Das show that tools like Woebot and Tess can meaningfully reduce symptoms of anxiety and depression [13, 1]. These results contribute to the wider acceptance of AI-supported self-help tools, although researchers such as Yu Hui Sim and Minerva caution that these systems can generate hallucinated or generic responses if not carefully supervised [31]. He and Yang Ni studied the broader role of large language models in counseling and mental-health assessment, they noted their potential and the ethical complexities involved both.

There is also a lot of writing about wearable biosensors. Vegesna compared ML models that use wearable data and he showed Recurrent Neural Networks achieve accuracy levels above 90%, and that hybrid CNN-RNN architectures or GAN-based augmentation strategies could further improve results [1]. A large review by Muetunda identifies Random Forest and SVM as the most widely used models, while the fuzzy inference systems also show impressive performance in several psychiatric classification tasks [19]. Other researchers, such as Montesines Nagayo and Gopalakrishnan, highlight complementary methodological needs, including robust data cleaning, self-supervised learning, and integration of fuzzy decision-support modules [33].

Personalization and multimodal systems are emerging as the next major frontiers. Macro-Micro Personalization Framework by Song et al. uses multimodal-multitask learning with advanced uncertainty handling mechanisms to tailor daily mental health predictions. Strategies for fusing multimodal data are discussed by Abinaya Gopalakrishnan such as early and delayed integration to overcome missing or inconsistent information, which is a common problem in real world monitoring where participants may not complete all questionnaires, or they do not wear sensors consistently [33].

Several reviews bring up ethical and methodological issues. While clinicians are increasingly recognizing ML’s usefulness, the small and homogeneous study samples limit generalizability very frequently as noted by Alimour [18]. Although monitoring focused studies tend to meet higher quality criteria, the intervention-oriented studies often fail to report basic metrics such as adherence and dropout rates as reported by Cruz-Gonzalez [23]. A mixed methods review by Gutierrez’s, which uses RoB 2 and GRADE, has reinforced the need for stronger reporting standards across biomedical AI research [26]. When more than 172,000 sources were screened by Okeibunor he argued that equitable and sustainable AI development is essential especially for low resource health systems [25].

We can see that AI is advancing progressively toward more integrated, multimodal, and therapeutic based usage. The AI systems are getting better at handling complicated behavioural and psychiatric data. These systems can perform things like bio-sensor monitoring, sentiment analysis, create personalized digital therapeutics, and even provide counseling to people using large language models. Researchers always talk about how vital it is to be able to understand, have ethical protections, work together across sectors, and be stringent with techniques. The merging of different types of data such as physiological, linguistic, behavioural, and contextual points to the future of mental health AI being in terms of precision, personalization, and openness. This growing body of research creates a strong basis for the development of safe, effective, and dependable AI-enhanced mental health care.

3 Methodology

We conducted this review in line with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The procedures below are described in enough detail to allow replication.

3.1 Database Search

The search was run on 22 September 2025 across eight databases: PubMed, PubMed Central, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Wiley Online Library, and a group of open-access publishers including Nature, PLOS ONE, JMIR, Frontiers, and MDPI. We also hand-screened proceedings from NeurIPS, ICML, ACL, EMNLP, and INTERSPEECH, since significant AI and ML work appears in conference venues before reaching journals. Search strings combined three term groups via Boolean operators: mental health condition terms (depression, anxiety, stress, suicide, schizophrenia, bipolar disorder); AI and ML method terms (machine learning, deep learning, transformer, BERT, RoBERTa, SVM, CNN, RNN, large language model); and data type descriptors (text, speech, EEG, ECG, HRV, electrodermal activity, wearable sensor).

3.2 Inclusion and Exclusion Criteria

We included a study if it met all six of the following conditions: (i) human participants with a recognized mental health condition; (ii) AI or ML applied to text, speech, or physiological data; (iii) quantitative evaluation metrics reported (accuracy, F1-score, sensitivity, specificity, or AUC-ROC); (iv) peer-reviewed journal or conference paper; (v) published January 2015 to September 2025; (vi) written in English. We excluded papers using only synthetic or acted data, animal studies, papers without usable metrics, grey literature, unreviewed preprints, and duplicates.

3.3 Screening and Data Extraction

Two reviewers were involved in a screening process of titles and abstracts where disputes were addressed through discussion. In situations where consensus was not arrived at then a third opinion was consulted. All records that came through this stage were then retrieved using full texts. In case of each selection study, we were able to extract: citation and year; target condition; modality and dataset; sample size and demographics where available; model architecture; evaluation metrics; whether external validation was performed; and reporting quality indicators such as confidence interval, calibration, cross-subject validation, and interpretability.

3.4 Quality Appraisal

Determining the quality was done based on criteria that were adapted to an NIH Quality Assessment Tool that was modified to fit AI diagnostic research. We assessed every study in four domains: sampling adequacy (sample size, diversity, transparency of recruitment); reference-standard appropriateness (diagnostic label validity and consistency); statistical rigor (train/test split treatment, confidence interval reporting); and reporting completeness (preprocessing, hyperparameters, limitations). A rating was given to each study as either Good, Fair, or Poor. Studies with low ratings were included in the review but handled with interpretive care all along.

3.5 Synthesis Approach

Formal meta-analysis was not appropriate given the heterogeneity of included studies in terms of conditions, modalities, datasets, and evaluation methods. We used narrative synthesis instead, grouping findings by modality, model type, clinical target, and quality rating. Where metrics were sufficiently comparable, we drew cross-modal comparisons and identified patterns of methodological weakness that cut across multiple studies.

4 Results

4.1 Study Selection

The database search returned 120 records. After removing duplicates, 100 papers went through title and abstract screening. Thirty-seven were excluded at this stage mainly for using synthetic data, missing evaluation metrics, or not being peer-reviewed. The remaining 63 met all our inclusion criteria and form the corpus for this review.

4.2 Study Characteristics

Publication dates ranged from 2015 to 2025, but output increased noticeably from 2020 onwards almost certainly reflecting the surge in mental health research triggered by the COVID-19 pandemic. Depression and MDD were the most studied conditions by a substantial margin. This reflects both their prevalence and the availability of labelled benchmark datasets, which we argue creates a bias in the research agenda that is worth naming explicitly. Stress and anxiety were also well-represented, particularly in wearable sensor studies. Suicidal ideation, schizophrenia, and bipolar disorder appeared less often but were present across several modalities.

4.3 Modality Distribution

Physiological signal studies formed the largest group — EEG, HRV, ECG, electrodermal activity, sleep and activity monitoring. Text-based NLP studies, mainly on social media and clinical text, were second. Speech studies came third. Multimodal studies were fewest in number overall but grew noticeably in frequency after 2022, suggesting the field is gradually moving in that direction.

4.4 Performance Findings by Modality

Accuracy results were reported to be high in all modalities at times dramatically so. The summarised peak performance on a per-modality basis is presented in Table 1. A structured breakdown of 12 representative studies as well as their major limitations is given in Table 2.

EEG-based deep learning models led the field in the physiological signals group, and some of the studies showed accuracies of 97–98.8% and F1-scores above 0.98. We wish to point out at this stage that these numbers must be handled with caution EEG research in our corpus had less than 30 subjects almost half of the time, and laboratory assumptions are not anything to do with actual EEG implementation. HRV and ECG models showed good performance as well, and CNN-LSTM hybrids always win in comparison with traditional classifiers. Dedgaonkar et al. generalized this to student populations where there was promising results [59].

Text-based transformer models, especially BERT and RoBERTa, reported accuracies of over 98% and F1-scores of up to 0.85 on depression classification and suicidal ideation recognition; SVM and logistic regression baselines were constantly outperformed. Both Khaliq and U and others show good text-based classification performance, with which there is, again, limited external validation [43, 55].

There was a greater variation in speech models. Depending on the task and dataset, accuracy reported ranged from about 75% to 99.82%. CNN-based and composite deep learning model-based treatments tended to perform better than conventional ML analysis of depressive state identification. The contributions of De Oliveira and others on anxiety detection were helpful work on clinical speech [13], whereas Jadhav and others studied emotion conscious systems with extraction of acoustic features [56]. Multimodal systems combining physiological and behavioural data normally achieved 90–94.58% accuracy and F1 of approximately 0.88.

These patterns are visualised in Figure 3 across seven model categories and four metrics: specificity, sensitivity, accuracy, and F1-score. A few findings stand out clearly. On all four metrics, EEG deep learning models are the first. BERT-based models outperform traditional text classifiers by a larger margin on sensitivity than on accuracy, which is a clinical issue: a false negative in a screening scenario carries real consequences. The poorest overall performance is indicated by speech traditional ML models. In all the categories, deep learning and transformer algorithms outperform classical algorithms, most evidently on recall-oriented measures. These figures, as has been mentioned above, belong to controlled experiments and real-world performance would certainly be lower.

Table 1. Modality vs. Peak Reported Performance

Modality	Typical Models	Peak Reported Performance	Representative Tasks
Physiological (EEG/HRV/ECG/EDA)	CNN, LSTM, CNN-LSTM, WS-BiLSTM-MA	97-98.8% accuracy; F1 > 0.98	Depression detection, stress classification
Text (NLP)	BERT, RoBERTa, Transfor-	>98% accuracy;	Suicidal ideation detection,

	mer ensembles	F1 up to 0.85	depression classification
Speech	CNNs, DNNs, hybrid DL	75-99.82% accuracy	Depression/anxiety screening; emotion recognition
Multimodal	CNN-LSTM, RNN ensembles, Fusion networks	90-94.58% accuracy; F1 $\approx$ 0.88	Combined physiological and behavioural assessment

Table 2. Representative Included Studies: Characteristics, Performance and Limitations.

Study	Condition	Modality	Model	Key Metric	Key Limitation	Relevance
Jiao [10]	Stress/Anxiety	Physiological (HRV, EDA)	SA-BFRNN	Precision: 89.23%	Single dataset, no external validation	Biosensor ML for continuous monitoring
Li [11]	Depression	Physiological (EEG)	RPDO-PSVM	Accuracy: 96%	Small sample, lab-controlled	Hybrid algorithm for EEG biomarker detection
Olatinwo [12]	Multiple	Physiological (IoT)	CNN-LSTM with LIME	F1 > 0.97	Limited demographic diversity	Interpretability in deep learning models
Zubair [14]	Suicidal Ideation	Text (social media)	LSTM/Transformer	F-measure: 0.83	Platform-specific limitations	NLP transfer learning for suicide risk
Baydili [27]	Depression/Suicide	Text (social media)	RoBERTa	Accuracy: >98%	Social media bias	Transformer superiority over classifiers
Allam [24]	Suicidal Ideation	Text	SVM/Transformer	High F1	Small dataset	SVM competitiveness alongside transformers
Monaco [7]	MDD	Multi-modal (Wearable)	ML ensemble	Protocol-based	No quantitative ML metrics	Ethical governance framework
Song [20]	Multiple	Multi-modal	Multimodal-Multitask	Mid-90% accuracy	Limited reproducibility	Personalised multi-modal prediction
Mu-etunda [19]	Multiple	Multiple	RF, SVM, Fuzzy	Varies	Heterogeneous study pool	RF/SVM dominance across modalities
Cruz-Gonzalez [23]	Multiple	Multiple	Systematic Review	Quality ratings	Lacks adherence metrics	Reporting gaps in AI interventions
Gutierrez [26]	Multiple	Multiple	Mixed Methods	GRADE/RoB 2	Limited quantitative synthesis	Need for standardised reporting



Fig. 1. Performance comparison of AI models across speech, text and physiological modalities.

4.5 Quality Assessment Findings

Our NIH-adapted appraisal rated roughly 59% of studies as good quality and 40% as Fair, with a small number rated Poor. The problems we saw most often: sample sizes too small for generalisation (acute in EEG and HRV research); no external validation; missing confidence intervals or calibration data; and incomplete reporting of preprocessing and hyperparameter choices. Fairness evaluations and subgroup analyses were almost entirely absent across all 63 studies, not just a subset. We consider this the single most significant gap in the literature, and we return to it in the discussion.

5 Discussion

5.1 The Depression Concentration Problem

The dominance of depression research in this corpus is striking. Part of this reflects genuine clinical importance, but a larger part, we suspect, reflects the availability of benchmark datasets like DAIC-WOZ and AVEC that make depression a convenient research target. The concern is circularity: more data on depression encourages more depression research, which further concentrates resources and attention on one condition at the expense of others. Bipolar disorder, schizophrenia, and PTSD are all clinically urgent, arguably harder to monitor, and much less well served by the current AI literature. Conditions with episodic symptom profiles or longer diagnostic timelines do not fit neatly into standard ML benchmarking workflows, and the field has not yet found good ways to address this.

5.2 Physiological Signals: Impressive Numbers, Fragile Foundations

The accuracy figures reported for EEG-based deep learning models are technically impressive some above 98%. We do not want to dismiss this work, but we think the numbers are being presented without enough context. Most of these studies used fewer than 30 participants, recruited from narrow demographic pools, and collected data under laboratory conditions with controlled environments and expert electrode placement. Under these conditions, a model can achieve high accuracy partly by learning participant-specific signal artefacts rather than genuine diagnostic patterns. Whether the same model would perform comparably on data from a new hospital, a different country, or patients with comorbidities is an open question that most papers do not even ask. Real-world EEG deployment also involves challenges motion artefacts, variable electrode placement, specialist equipment requirements that are absent from controlled laboratory settings.

5.3 Transformer NLP: Strong Performance, Complicated Picture

BERT and RoBERTa have genuinely transformed what is possible in text-based mental health detection, and the accuracy figures across multiple studies reflect real technical progress. The harder question is what those figures actually mean for clinical application. Social media datasets carry well-known demographic biases: they skew young, urban, and English-speaking, and the relationship between what someone posts online and what a clinician would observe in a consultation is not straightforward. A model achieving 98% on Twitter data might perform quite differently on clinical transcripts from a rural

hospital in another country. There are also ethical complications that accuracy numbers do not address: consent, data ownership, and the risks of automated mental health surveillance at population scale are not resolved by a good F1-score.

5.4 Speech Models: The Ecological Validity Gap

Speech-based models have shown consistent predictive value, and features like pitch variability, pausing, and vocal energy have held up across multiple studies [50]. The issue we keep coming back to is ecological validity. A substantial proportion of speech datasets in this literature were elicited under controlled conditions: participants reading passages aloud or responding to scripted prompts. This is a poor substitute for the spontaneous, contextually embedded speech that happens in real clinical encounters or everyday conversation. Whether the acoustic signatures of depression identified in scripted recordings transfer to naturalistic speech is a question that has not been adequately investigated, and it is one that would need to be answered before these tools could be trusted in clinical screening.

5.5 Multimodal Fusion: Promising in Theory, Underspecified in Practice

The consistent superiority of multimodal systems over single-modality approaches is one of the clearest patterns in this review, and it makes sense conceptually: mental health conditions manifest across biological, linguistic, and behavioural dimensions simultaneously, and combining signals from different sources should, in principle, produce a richer and more robust picture. The practical gap, however, is significant. Nearly all multimodal studies in our corpus assume complete data availability all sensor streams present and intact for every participant at every time point. In real deployments, this assumption fails constantly. How well current multimodal architectures handle missing modalities has barely been studied, and it is one of the most practically important questions the field needs to address.

5.6 Recurring Methodological Problems

Looking across the full corpus, the same methodological problems surface repeatedly. The most consequential is the absence of external validation. Training and testing on the same dataset — or on closely related datasets from the same lab — is not a valid test of generalisability. It tells you how well a model fits a particular data distribution, not whether it will work on new populations or in new settings. Yet this is precisely the kind of evidence that would be needed before any responsible clinical deployment, and it is almost entirely absent from the literature we reviewed.

The reporting practices contribute to the problem: confidence intervals, calibration data, and error analysis is mentioned infrequently, and often description of the used pipelines to preprocess is too vague to be replicated. The fact that less than 50 percent of explainable AI tools are used, i.e., LIME, SHAP, attention visualisation are used is also a related issue. A system to which clinicians cannot pose queries cannot be trusted by clinicians in critical situations such as suicide risk evaluation.

5.7 Fairness and Bias: The Unrecognized Issues

Among all the loopholes that we discovered, the fact that we found almost zero fairness assessment is what bothers us the most. Training data systematically excluding or underrepresenting specific groups, as it is not the exception but the norm in this literature, results in models that can perform well in some patients and poorly in others. This may only worsen the current health disparities instead of resolving them should it be released into clinical practice [8, 37, 58]. This is an unresolved critical issue that is identified by Gerantia in his review of AI in psychiatry and by Terra and colleagues in his ethical analysis [37, 58]. It is not a hypothetical issue because the dissimilarity between demographic subgroups in terms of the performance of AI has been extensively reported in other medical areas. The stakes of the misidentification are particularly high in mental health where marginalisation and low-income groups bear the disproportionate burden of mental health issues. Almost all the studies that we examined reported subgroup performance analysis, and many of them did not even provide a demographic description of their dataset, which is essential to judge representativeness.

5.8 What Clinical Translation Would Actually Require

The image generated by this review is of a field that has actually achieved technical advances, but still needs to undertake the more difficult task of getting ready to be clinically deployed. Sealing such a gap would take both external and longitudinal validation on diverse populations and settings; more representative data and architectural fairness assessment; interpretable architectures that can be actually used by clinicians; and ethical governance structures that acknowledge the sensitivities of mental health data. Sim and Choo's vision of an AI-enhanced mental health ecosystem built around human-AI collaboration rather than AI as a replacement for clinical judgment seems to us the most credible near-term direction [25]. Tools that support

clinicians — flagging risk, surfacing patterns, prompting follow-up — without bypassing them are more likely to earn trust, more likely to be used, and more likely to catch their own errors.

6 Limitations and Future Directions

6.1 Limitations of This Review

It is what this review cannot do we should be clear. The search was only made in English on a single date and hence the issues in other languages or places which we do not search will have been overlooked. The heterogeneity of included studies, which, cross cutting conditions, modalities, datasets and evaluation frameworks, precluded any formal meta-analysis; narrative synthesis is more interpretively flexible by definition. Our quality appraisal tool, made based on existing tools, was not AI-based mental health research-specific and validated. And the field proceeds at leaps and bounds: it may well seem that much work has been done since our September 2025 deadline.

6.2 Issues That Should Be Addressed Across the Field

In addition to the limitation of our own review, there are a number of structural issues that restrict the evidence base at present. The samples are too small, especially in the field of physiological signals. There is too narrow condition coverage. The problem of methodological inconsistency between studies complicates the process of synthesis and complicates the process of replication. And even with the best intentions, the lack of common reporting standards implies that even good papers tend to omit details which would be required to make sense of or extend their findings.

6.3 Directions for Future Work

Some of the directions appear especially significant. Multi-site longitudinal validation studies are sorely needed not more single site experiments with new model architectures. Similar reporting criteria, like CONSORT or STARD, should be formulated and implemented in the community that requires adequate disclosure of validation procedures, demographic factors, and fairness testing. Explainable AI must cease being optional feature and become a design requirement particularly with high-stakes applications. Combination with clinical processes and electronic health records is a serious topic of research, including the governance issues embedded in it. Federated learning and difference privacy are privacy-preserving methods that ought to be embraced more due to the sensitivity of mental health information. And the discipline must make the deliberate attempt to diversify the subjects of its inquiry beyond the present depression-laden preoccupation.scale.

7 Conclusions

Assuming the existence of a single truthful at a glance overview of the current state of AI-based mental health monitoring, it would be accurate to state that it is technically impressive, but clinically premature. Deep neural network models of EEG, transformer neural language systems, speech synthesizers, and multimodal integrative models have all delivered results that are truly impressive. However, the overwhelming preponderance of these findings are of small, controlled studies that are little validated, their datasets limited and practically no consideration of fairness or interpretability. It is not well known whether they would stand in a real clinical setting, when they are applied to a large group of patients, and by clinicians who must know and be able to give explanations of the system outputs.

This does not imply that the work is worthless and the direction is erroneous. It implies that the discipline must become more adult in certain significant aspects. A high achievement on a benchmark is not the final remark. The procedures between this point and the responsible clinical deployment such as external validation, demographic representativeness, interpretability, ethical governance are not merely post hoc. They are the actual hard work.

We would recommend this in a simple manner. The appropriate near-term model is the use of hybrid human-AI systems that assist clinicians but do not take their place. The community that conducts research must take validation, diversity and explainability seriously instead of attempting to optimize accuracy on increasingly familiar benchmarks. There is potential in the technology. What has to be done is the realisation of it responsibly.

References

1. Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024b). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine Surgery and Public Health*, 3, 100099. <https://doi.org/10.1016/j.glmedi.2024.100099>
2. International Research Journal of Computer Science. (2020). *International Research Journal of Computer Science*. <https://doi.org/10.26562/irjcs>
3. Shimada, K. (2023). The Role of Artificial Intelligence in Mental Health: A review. *Science Insights*, 43(5), 1119–1127. <https://doi.org/10.15354/si.23.re820>
4. Das, K. P., & Gavade, P. (2024). A review on the efficacy of artificial intelligence for managing anxiety disorders. *Frontiers in Artificial Intelligence*, 7, 1435895. <https://doi.org/10.3389/frai.2024.1435895>
5. Kargarandehkordi, A., Li, S., Lin, K., Phillips, K. T., Benzo, R. M., & Washington, P. (2025). Fusing Wearable Biosensors with Artificial Intelligence for Mental Health Monitoring: A Systematic Review. *Biosensors*, 15(4), 202. <https://doi.org/10.3390/bios15040202>
6. Long, N., Lei, Y., Peng, L., Xu, P., & Mao, P. (2022). A scoping review on monitoring mental health using smart wearable devices. *Mathematical Biosciences & Engineering*, 19(8), 7899–7919. <https://doi.org/10.3934/mbe.2022369>
7. Monaco, F., Vignapiano, A., Piacente, M., Farina, F., Pagano, C., Marenni, A., Leo, S., Vecchi, C., Mancuso, C., Prisco, V., Iodice, D., Auricchio, A., Cavaliere, R., D'Agosto, A., Fornaro, M., Solmi, M., Corrivetti, G., & Fasano, A. (2024). Innova4Health: an integrated approach for prevention of recurrence and personalized treatment of Major Depressive Disorder. *Frontiers in Artificial Intelligence*, 7, 1366055. <https://doi.org/10.3389/frai.2024.1366055>
8. Kang, M., & Chai, K. (2022). Wearable sensing systems for monitoring mental health. *Sensors*, 22(3), 994. <https://doi.org/10.3390/s22030994>
9. Liu, L., Liu, L., Wafa, H. A., Tydeman, F., Xie, W., & Wang, Y. (2024). Diagnostic accuracy of deep learning using speech samples in depression: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, 31(10), 2394–2404.
10. Jiao, C. (2024). A mental health monitoring system based on intelligent algorithms and biosensors: Algorithm behaviour analysis and intervention strategies. *Molecular & Cellular Biomechanics*, 21(1), 357. <https://doi.org/10.62617/mcb.v21i1.357>
11. Li, Y. (2024). Research on the construction of biosensor-assisted mental health monitoring system and talent training strategy. *Molecular & Cellular Biomechanics*, 21(4), 814. <https://doi.org/10.62617/mcb814>
12. Olatinwo, D., Abu-Mahfouz, A., & Myburgh, H. (2025). Mental Disorder Assessment in IoT-Enabled WBAN Systems with Dimensionality Reduction and Deep Learning. *Journal of Sensor and Actuator Networks*, 14(3), 49. <https://doi.org/10.3390/jsan14030049>
13. Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., ... & Yaniv-Rosenfeld, A. (2019). Big data analytics and AI in mental healthcare. *arXiv preprint arXiv:1903.12071*.
14. Zuberi, A. H., Anees, A., Anjum, N., Warsi, A. H., Khan, P. R., Singh, S. K., ... & Ranjan, R. (2025). Machine Learning-Based Sentiment Analysis for Suicide Prevention and Mental Health Monitoring in Educational Institutions. *Journal of Neonatal Surgery*, 14(5s).
15. Shaik, T., Tao, X., Higgins, N., Xie, H., Gururajan, R., & Zhou, X. (2022, October). AI enabled RPM for mental health facility. In *Proceedings of the 1st ACM Workshop on Mobile and Wireless Sensing for Smart Healthcare* (pp. 26–32).
16. He, T., Fu, G., Yu, Y., Wang, F., Li, J., Zhao, Q., ... & Yang, B. X. (2023). Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *arXiv preprint arXiv:2312.04578*.
17. Pandey, H. M. (2025). Artificial Intelligence in Mental Health and Well-Being: Evolution, current applications, future challenges, and Emerging evidence. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2501.10374>
18. Alimour, S. A., Alnono, E., Aljasm, S., Farran, H. E., Alqawasm, A. A., Alrabeei, M. M., Shwede, F., & Aburayya, A. (2024). The quality traits of artificial intelligence operations in predicting mental healthcare professionals' perceptions: A case study in the psychotherapy division. *Journal of Autonomous Intelligence*, 7(4). <https://doi.org/10.32629/jai.v7i4.1438>
19. Muetunda, F., Sabry, S., Jamil, M. L., Pais, S., Dias, G., & Cordeiro, J. (2024). AI-Assisted Diagnosing, Monitoring and Treatment of Mental Disorders: A survey. *ACM Transactions on Computing for Healthcare*, 5(4), 1–24. <https://doi.org/10.1145/3681794>
20. Song, M., Yang, Z., Triantafyllopoulos, A., Zhang, Z., Nan, Z., Tang, M., Takeuchi, H., Nakamura, T., Kishi, A., Ishizawa, T., Yoshiuchi, K., Schuller, B., & Yamamoto, Y. (2024). Empowering Mental Health Monitoring: Macro-Micro Personalization Framework for Multimodal-Multitask Learning (PrEPrInt). *JMIR Mental Health*, 11, e59512. <https://doi.org/10.2196/59512>
21. Omiyefa, S. (2025). Artificial intelligence and machine learning in precision mental health diagnostics and predictive treatment models. *International Journal of Research Publication and Reviews*, 6(3), 85–99. <https://doi.org/10.55248/gengpi.6.0325.1107>
22. Bhat, A. (2025). An AI-Enabled framework for mood analysis and mental health support using user interaction data. *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 09(07), 1–9. <https://doi.org/10.55041/ijsem51154>
23. Cruz-Gonzalez, P., He, A. W., Lam, E. P., Ng, I. M. C., Li, M. W., Hou, R., Chan, J. N., Sahni, Y., Guasch, N. V., Miller, T., Lau, B. W., & Vidana, D. I. S. (2025). Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. *Psychological Medicine*, 55, e18. <https://doi.org/10.1017/s0033291724003295>
24. Allam, H., Davison, C., Kalota, F., Lazaros, E., & Hua, D. (2025). AI-Driven Mental Health Surveillance: Identifying suicidal ideation through machine learning techniques. *Big Data and Cognitive Computing*, 9(1), 16. <https://doi.org/10.3390/bdcc9010016>
25. Okeibunor, J. C., Jaka, A., Iwu-Jaja, C. J., Idemili-Aronu, N., Ba, H., Zantsi, Z. P., Ndlambe, A. M., Mavundza, E., Muneene, D., Wiysonge, C. S., & Makubalo, L. (2023). The use of artificial intelligence for delivery of essential health services across WHO regions: a scoping review. *Frontiers in Public Health*, 11, 1102185. <https://doi.org/10.3389/fpubh.2023.1102185>
26. Gutierrez, G., Stephenson, C., Eadie, J., Asadpour, K., & Alavi, N. (2024b). Examining the role of AI technology in online mental healthcare: opportunities, challenges, and implications, a mixed-methods review. *Frontiers in Psychiatry*, 15, 1356773. <https://doi.org/10.3389/fpsy.2024.1356773>

27. Baydili, İ., Tasci, B., & Tasci, G. (2025). Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection. *Behavioral Sciences*, 15(3), 352. <https://doi.org/10.3390/bs15030352>
28. Sousa, R. T., Silva, G. D. M., Pinto, P. L. L., Backes, J., Mota, A. S., Paixão, T. M., Teixeira, M. G. S., Hisatugu, W. H., Garcia, A. S., Prado, K. S., Dias, R. S., Galletta, M. A., & Tavares, H. (2025). Depression and anxiety screening for pregnant women via free conversational speech in naturalistic condition. *IEEE Access*, 13, 52149–52156. <https://doi.org/10.1109/access.2025.3552659>
29. Baydili, İ., Tasci, B., & Tasci, G. (2025). Artificial Intelligence in Psychiatry: A review of Biological and Behavioral data analyses. *Diagnostics*, 15(4), 434. <https://doi.org/10.3390/diagnostics15040434>
30. Karimzadeh, D., & Saeedi, A. (2025). AI for Mental Health Assessment and Intervention: A Systematic Review. *International Journal of Modern Achievement in Science, Engineering and Technology.*, 2(1), 96–104. <https://doi.org/10.63053/ijset.65>
31. Minerva, F., & Giubilini, A. (2023). Is AI the future of mental healthcare? *Topoi*, 42(3), 809–817. <https://doi.org/10.1007/s11245-023-09932-3>
32. Cosic, K., Kopilas, V., & Jovanovic, T. (2024). War, emotions, mental health, and artificial intelligence. *Frontiers in Psychology*, 15, 1394045. <https://doi.org/10.3389/fpsyg.2024.1394045>
33. Gopalakrishnan, A., Gururajan, R., Zhou, X., Venkataraman, R., Chan, K. C., & Higgins, N. (2024). A survey of autonomous monitoring systems in mental health. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 14(3). <https://doi.org/10.1002/widm.1527>