International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)

ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026

www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

[Artificial Intelligence in the Detection and Diagnosis of Depression and Anxiety: A Narrative Review of Data Modalities, Machine Learning Methods and Translational Challenges]

***Mansi*[*1], *Dr. Grace M. BasanShrieh*[*2]**

[*1] *Phd Scholar, Department of Psychology, William Caery University, Meghalaya, India*

[*2] *Assistant Professor and Guide, Department of Psychology, William Caery University, Meghalaya, India*

ABSTRACT

Depression and anxiety are among the most prevalent mental disorders worldwide, yet a large proportion of affected individuals remain unrecognised in routine care. Artificial intelligence (AI) has been proposed as a way to support earlier and more consistent identification by learning patterns from questionnaires, electronic health records, language, speech and passively sensed behaviour. This narrative review synthesises the evidence on AI-assisted detection and diagnosis of depression and anxiety across five domains: clinical foundations of assessment, data modalities, machine learning algorithms, evaluation practice, and ethical and governance requirements. Literature was identified in PubMed, Scopus, IEEE Xplore, Web of Science and Google Scholar and synthesised thematically. The review finds that standard instruments such as the PHQ-9 and GAD-7 offer sensitivity and specificity in the range of 0.82–0.89 at conventional cut-offs, whereas unassisted clinical recognition of depression in primary care detects only about half of cases, which defines the gap AI tools aim to close. Classical models (logistic regression, support vector machines, random forests, gradient boosting) remain competitive on structured data, while deep learning and transformer-based language models dominate unstructured text and speech. Across modalities, however, the evidence base is limited by small and homogeneous samples, reliance on self-report proxies as ground truth, scarce external validation, rarely reported calibration, and little fairness auditing. The review concludes that AI should be positioned as clinician-supervised decision support rather than an autonomous diagnostic agent, and it sets out a research agenda built around multimodal data, external and temporal validation, calibrated and explainable outputs, subgroup fairness, and reporting under the TRIPOD+AI guideline.

KEYWORDS: *Artificial intelligence; machine learning; depression; anxiety; mental health screening; natural language processing; speech analysis; digital phenotyping; explainable AI; algorithmic fairness*.

Received	Accepted	Published	Article Type
18/07/2026	20/08/2026	25/09/2026	Research Paper

© 20XX IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)

ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026

www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

***Abstract*—** Depression and anxiety are among the most prevalent mental disorders worldwide, yet a large proportion of affected individuals remain unrecognised in routine care. Artificial intelligence (AI) has been proposed as a way to support earlier and more consistent identification by learning patterns from questionnaires, electronic health records, language, speech and passively sensed behaviour. This narrative review synthesises the evidence on AI-assisted detection and diagnosis of depression and anxiety across five domains: clinical foundations of assessment, data modalities, machine learning algorithms, evaluation practice, and ethical and governance requirements. Literature was identified in PubMed, Scopus, IEEE Xplore, Web of Science and Google Scholar and synthesised thematically. The review finds that standard instruments such as the PHQ-9 and GAD-7 offer sensitivity and specificity in the range of 0.82–0.89 at conventional cut-offs, whereas unassisted clinical recognition of depression in primary care detects only about half of cases, which defines the gap AI tools aim to close. Classical models (logistic regression, support vector machines, random forests, gradient boosting) remain competitive on structured data, while deep learning and transformer-based language models dominate unstructured text and speech. Across modalities, however, the evidence base is limited by small and homogeneous samples, reliance on self-report proxies as ground truth, scarce external validation, rarely reported calibration, and little fairness auditing. The review concludes that AI should be positioned as clinician-supervised decision support rather than an autonomous diagnostic agent, and it sets out a research agenda built around multimodal data, external and temporal validation, calibrated and explainable outputs, subgroup fairness, and reporting under the TRIPOD+AI guideline.

***Keywords*—** *Artificial intelligence; machine learning; depression; anxiety; mental health screening; natural language processing; speech analysis; digital phenotyping; explainable AI; algorithmic fairness*

1. INTRODUCTION

1.1 Background and Burden

Depression and anxiety are common, often recurrent conditions that affect mood, cognition, behaviour, sleep, relationships and occupational functioning. The Global Burden of Disease Study 2019 estimated roughly 280 million prevalent cases of depressive disorders and 301 million cases of anxiety disorders worldwide, placing both among the leading causes of years lived with disability [1]. The COVID-19 pandemic added an estimated 53.2 million cases of major depressive disorder and 76.2 million cases of anxiety disorders in 2020 alone, increases of 27.6% and 25.6%, respectively [2]. Fig. 1 summarises these figures. The Lancet Commission on global mental health has stressed that this burden falls heavily on low- and middle-income settings, where specialist services are scarce and treatment gaps are widest [3].

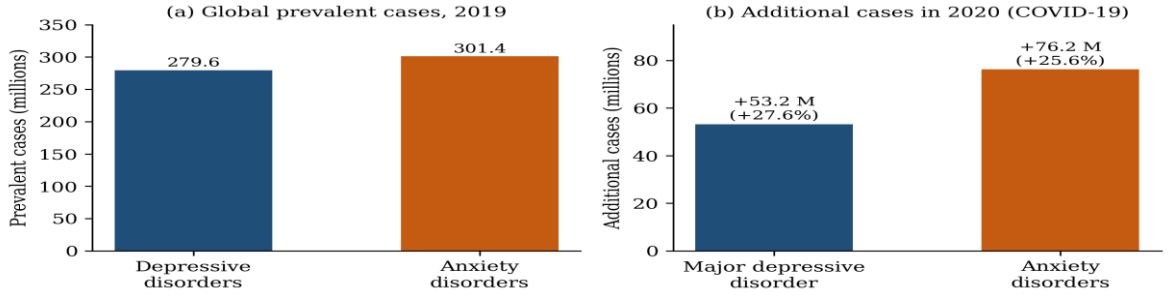


Fig. 1. Global burden of depressive and anxiety disorders. (a) Prevalent cases in 2019 from the Global Burden of Disease Study [1]. (b) Additional cases estimated for 2020 as a consequence of the COVID-19 pandemic [2].

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)
ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026
www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

The two conditions frequently co-occur. Epidemiological surveys show extensive comorbidity between mood and anxiety disorders [4], and co-occurring anxiety is associated with greater severity, poorer treatment response and higher chronicity of depression [5]. Any assessment approach, human or computational, therefore has to handle overlapping and dimensional symptom presentations rather than two cleanly separated categories.

1.2 The Diagnostic Gap

Diagnosis depends on clinical interview, history and professional judgement, supported by standardised instruments. In busy primary care settings this process misses many people. A large meta-analysis found that general practitioners correctly identified depression in about half of affected patients (sensitivity 50.1%, specificity 81.3%) [6]. Stigma, limited consultation time, somatic presentations and the difficulty some people have in describing their experiences all contribute to under-recognition. Brief self-report tools help, but they are not diagnostic on their own, and their positive predictive value falls steeply when prevalence is low.

Artificial intelligence (AI), and machine learning (ML) in particular, has been proposed as a way to narrow this gap. ML models can combine many weak signals, such as questionnaire items, clinical codes, word use, vocal prosody and daily activity, into a single risk estimate, and they can be applied consistently at scale [7], [8]. Recent reviews describe rapid growth in AI studies of depression and anxiety [9], [10], the emergence of large language models (LLMs) in psychiatry [11], and parallel work on related outcomes such as suicide risk [12]. These reviews also point to recurring methodological weaknesses that limit clinical translation [13], [14].

1.3 Objectives of the Review

Drawing on the doctoral work *AI Assisted Diagnosis of Depression and Anxiety*, this review had four objectives: (i) to describe the clinical assessment foundations against which AI tools must be judged; (ii) to map the data modalities and ML methods used for detecting depression and anxiety; (iii) to examine evaluation practice, including validation, calibration and explainability; and (iv) to identify ethical, fairness and governance requirements and the research gaps that stand between current models and responsible clinical use.

2. REVIEW METHODOLOGY

2.1 Search Strategy

A structured narrative review was conducted. PubMed, Scopus, IEEE Xplore, Web of Science and Google Scholar were searched for peer-reviewed work published between 2000 and 2025, with seminal methodological papers included regardless of date. Search strings combined three concept blocks (Table 1) using Boolean operators, and reference lists of key reviews were hand-searched. The pathway is summarised in Fig. 2.

Table 1. Concept blocks used to construct the literature search

Concept block	Representative search terms
Condition	depression, depressive disorder, major depressive disorder, anxiety, generalised anxiety disorder, common mental disorders
Technology	artificial intelligence, machine learning, deep learning, natural language processing, speech analysis, digital phenotyping, large language model
Task and outcome	detection, screening, diagnosis, classification, prediction, risk estimation, validation, explainability, fairness

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)

ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026

www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

Search window: 2000–2025; English language; peer-reviewed sources

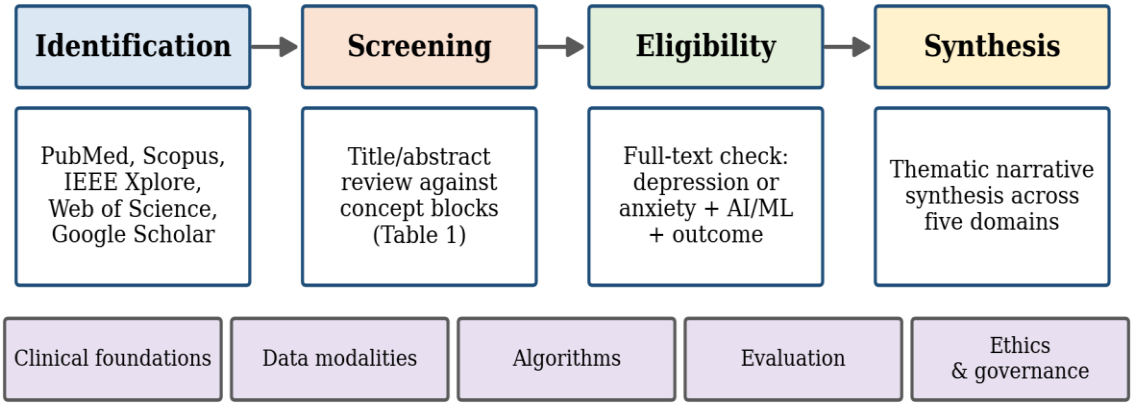


Fig. 2. Review pathway, from database identification through screening and eligibility assessment to thematic synthesis across five domains.

2.2 Eligibility and Synthesis

Studies were eligible if they addressed depression or anxiety, applied an AI or ML method for detection, screening, diagnosis or risk estimation, and reported a quantitative performance outcome or a substantive methodological, ethical or governance analysis. Opinion pieces without methodological content and studies of unrelated disorders were excluded. Because the included work was heterogeneous in population, modality, reference standard and metric, a meta-analysis was not appropriate. Findings were instead organised thematically into the five domains shown in Fig. 2.

3. CLINICAL FOUNDATIONS OF ASSESSMENT

3.1 Symptom Dimensions

Depression is characterised by persistent low mood or loss of interest together with changes in sleep, appetite, energy, concentration and self-worth, and in severe cases thoughts of death. Anxiety disorders involve excessive worry, restlessness, autonomic arousal and avoidance. Both conditions show cognitive, emotional, behavioural and physiological expressions, which is why diverse digital signals can carry information about them. The same diversity makes ground-truth labelling difficult: the symptom profiles of two people who meet criteria for the same disorder can differ markedly.

3.2 Standardised Instruments as Benchmarks

The Patient Health Questionnaire-9 (PHQ-9) and the Generalised Anxiety Disorder-7 (GAD-7) are the most widely used brief measures and the most common labels in AI studies. In the original validation, a PHQ-9 score of 10 or more had a sensitivity and specificity of 88% for major depression [15]. An individual participant data meta-analysis against semi-structured interviews confirmed a sensitivity of 0.88 and specificity of 0.85 at the same cut-off [16]. For the GAD-7, a cut-off of 10 gave a sensitivity of 89% and specificity of 82% for generalised anxiety disorder [17], and a later diagnostic meta-analysis supported acceptable accuracy across nearby cut-offs [18]. Fig. 3 compares these benchmarks with unassisted clinical recognition.

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)

ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026

www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

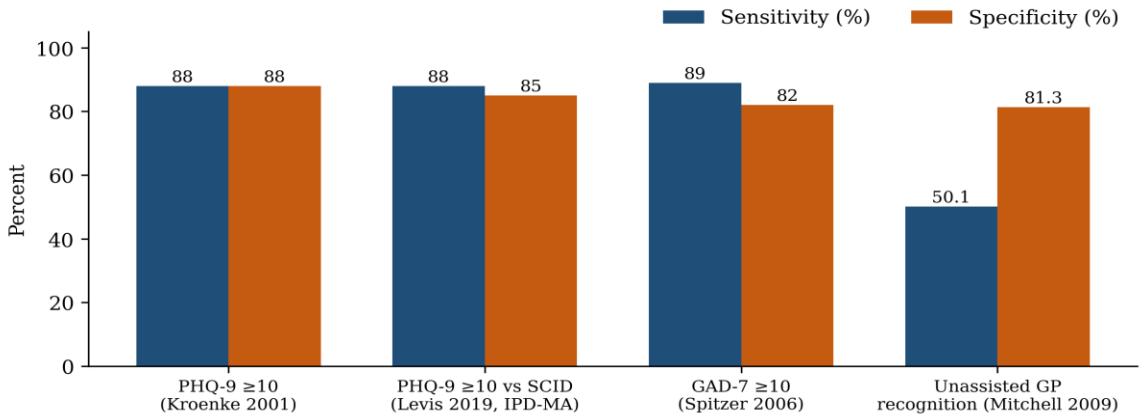


Fig. 3. Benchmark accuracy of standard screening instruments compared with unassisted general-practitioner recognition of depression. Values from [15], [16], [17] and [6].

These figures matter for AI research in two ways. First, they set a practical bar: an AI model that only reproduces PHQ-9 accuracy from far more intrusive data adds little. Second, when questionnaire scores are used as labels, model performance is capped by the error in the label itself. A model trained to predict PHQ-9 scores learns the questionnaire, not the diagnosis. Structured clinical interviews therefore remain the preferred reference standard wherever they are feasible.

4. DATA MODALITIES FOR AI-ASSISTED DETECTION

Table 2 summarises the main data sources, the features typically derived from them, and their principal strengths and limitations. Fig. 4 shows how they can be combined in a multimodal architecture.

Table 2. Data modalities used in AI-assisted detection of depression and anxiety

Modality	Typical features	Strengths	Limitations
Questionnaires	Item responses, total and sub-scale scores, response timing	Validated, low cost, directly interpretable	Self-report bias; label circularity when used as both input and outcome
Electronic health records	Diagnoses, prescriptions, utilisation, clinical notes	Large longitudinal samples; real-world care	Coding inconsistency; missing data; privacy constraints
Language and social media	Lexical, affective, pronoun use, topics, contextual embeddings	Rich, naturalistic, scalable	Platform and cultural bias; consent; non-clinical labels
Speech	Prosody, pitch, pause, energy, spectral and voice-quality features	Non-invasive; captures psychomotor change	Recording conditions, language and accent effects
Wearables and smartphones	Mobility, GPS variance, sleep, heart rate, phone use	Continuous, passive, objective	Adherence, device heterogeneity, battery and privacy issues

4.1 Questionnaires and Electronic Health Records

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)
ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026
www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

Item-level questionnaire data are the simplest input. ML can weight items, detect response patterns and shorten instruments adaptively. Electronic health records (EHRs) provide larger, longitudinal samples in which prior diagnoses, prescriptions, service use and free-text notes can be modelled. They also carry coding inconsistencies, missing data and institution-specific practices that can be learned as spurious shortcuts [14].

4.2 Language and Social Media

Language reflects cognitive and affective states. Reviews of social media studies report consistent associations between depression and greater use of first-person singular pronouns, negative emotion words and themes of loneliness and hopelessness [19]. In a study linking consented Facebook posts to medical records, language predicted a later clinical diagnosis of depression with moderate accuracy, comparable to screening questionnaires [20]. Transformer models such as BERT [21] capture context better than word-count approaches, and LLMs extend this further. LLMs also raise new concerns about hallucination, bias and privacy [11].

4.3 Speech and Vocal Biomarkers

Depression is associated with reduced prosodic variability, slower speech rate, longer pauses and changes in voice quality, which are consistent with psychomotor retardation [22]. A systematic review of automated speech assessment across psychiatric disorders found promising accuracy but noted small samples, heterogeneous feature sets and a lack of cross-corpus validation [23]. Recording environment, language and accent strongly affect acoustic features, which makes generalisation a central concern.

4.4 Wearables and Smartphone Sensing

Personal sensing, also called digital phenotyping, uses smartphones and wearables to capture mobility, sleep, physical activity, social interaction and phone use passively [24]. In an early exploratory study, GPS-derived features such as location variance and circadian regularity of movement were associated with PHQ-9 severity [25]. The digital psychiatry literature describes the potential of such data for continuous monitoring, and also persistent challenges of engagement, device heterogeneity and replication [26].

4.5 Multimodal Fusion

No single modality captures the full clinical picture, so multimodal fusion is an active research direction. Early fusion concatenates features before modelling, late fusion combines modality-specific predictions, and hybrid schemes learn joint representations [27]. Fusion can improve robustness when one channel is noisy or missing. It also increases model complexity, data requirements and the risk of overfitting, and it makes explanation harder.

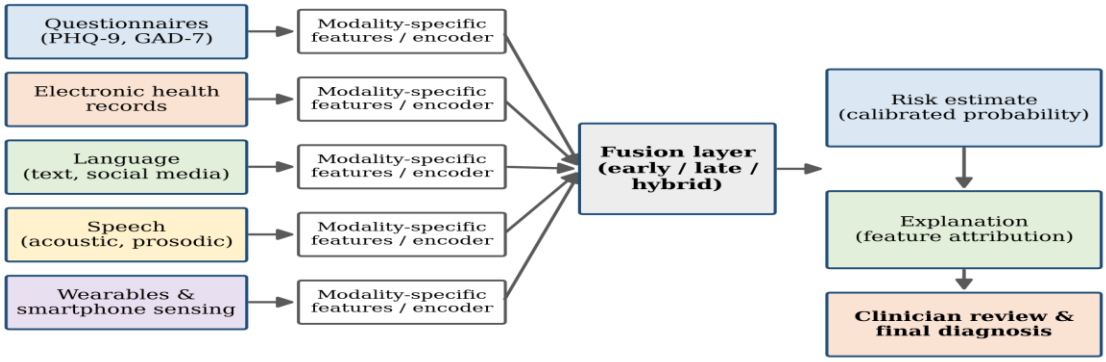


Fig. 4. Generic multimodal architecture for AI-assisted assessment. Modality-specific features are fused into a calibrated risk estimate that is accompanied by an explanation and reviewed by a clinician, who makes the diagnosis.

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)
ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026
www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

5. MACHINE LEARNING METHODS

5.1 Classical Supervised Learning

Logistic regression remains the reference baseline because it is transparent and well calibrated on structured data. Support vector machines find maximum-margin decision boundaries and handle high-dimensional feature spaces through kernels [28]. Random forests average many decorrelated decision trees and capture non-linear interactions [29]. Gradient boosting builds trees sequentially to correct residual errors and often performs best on tabular data [30]. For moderate-sized clinical datasets, several reviews report that these methods perform comparably to more complex models [8], [13]. The choice between them should rest on data structure, sample size and interpretability requirements rather than on novelty [31].

5.2 Deep Learning and Language Models

Deep neural networks learn hierarchical representations directly from raw or minimally processed data [32]. Convolutional networks suit spectrograms and images. Recurrent networks such as long short-term memory (LSTM) models capture temporal dependencies in speech, text and sensor streams [33]. Transformers, pretrained on large corpora and fine-tuned for mental health tasks [21], now dominate text-based detection. These gains come at the cost of large data requirements, computational expense and limited intrinsic interpretability, a trade-off shown qualitatively in Fig. 5 and summarised in Table 3.

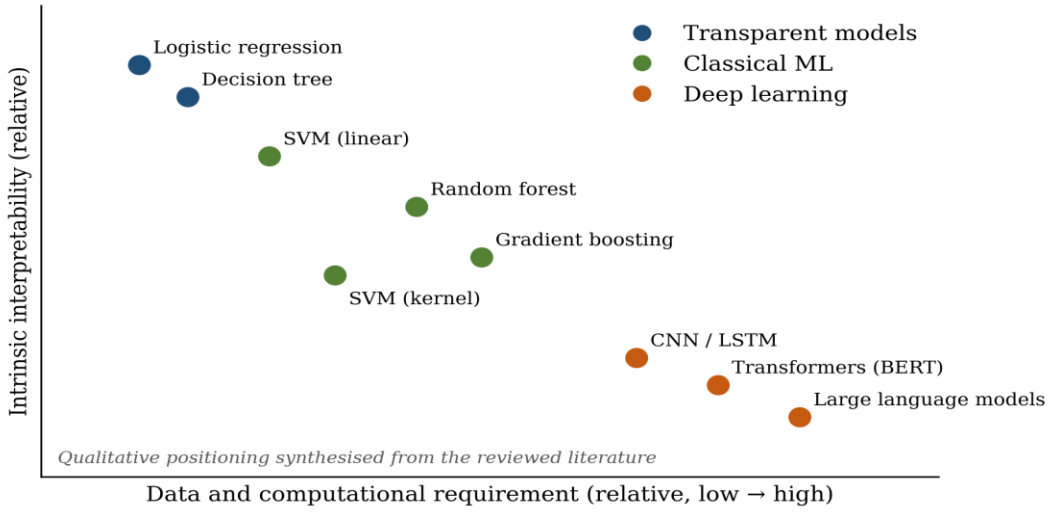


Fig. 5. Qualitative positioning of common algorithm families by intrinsic interpretability and by data and computational requirements, synthesised from the reviewed literature.

Table 3. Machine learning algorithm families applied to depression and anxiety detection

Algorithm	Typical data	Advantages	Cautions
Logistic regression	Questionnaire items, EHR variables	Interpretable coefficients; good calibration	Linear unless interactions are specified
Support vector machine [28]	High-dimensional text or acoustic features	Effective with moderate samples	Kernel choice; probability outputs need calibration
Random forest [29]	Mixed tabular data	Captures interactions; robust to noise	Less transparent; can be poorly calibrated

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)
ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026
www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

Gradient boosting [30]	Tabular EHR and sensor summaries	Often strongest on tabular data	Overfitting without careful tuning
CNN / LSTM [33]	Spectrograms, sequences, sensor streams	Learns temporal and spectral patterns	Needs large data; opaque
Transformers / LLMs [11], [21]	Clinical notes, transcripts, posts	Contextual language understanding	Data privacy, hallucination, bias, cost

5.3 Handling Class Imbalance

Clinical datasets are usually imbalanced, with fewer positive cases than negative ones. Resampling methods such as SMOTE [34], class weighting and threshold adjustment are common remedies. Resampling must be applied only within training folds, because applying it before splitting leaks information and inflates performance estimates.

6. EVALUATION AND VALIDATION PRACTICE

6.1 Performance Metrics

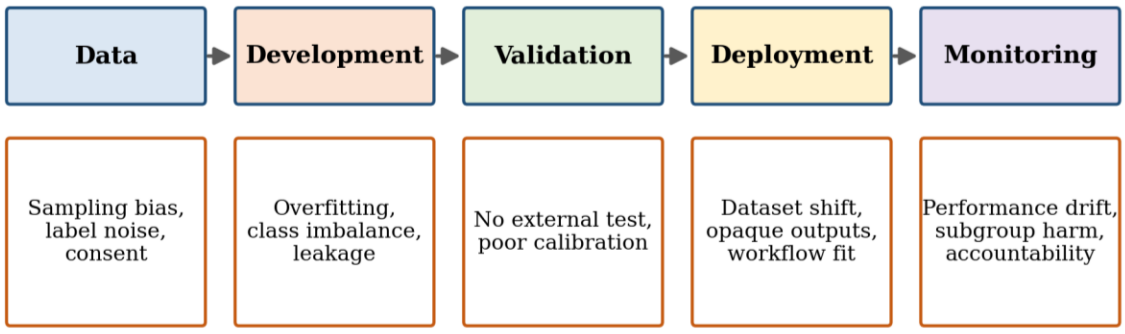
Accuracy is misleading when classes are imbalanced. Sensitivity, specificity, positive and negative predictive values, F1 score, the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC) give a fuller picture. PR-AUC is particularly informative in imbalanced settings [35]. Because predictive values depend on prevalence, performance reported in enriched research samples can overstate usefulness in the general population.

6.2 Validation, Calibration and Clinical Utility

Internal validation through cross-validation estimates performance on unseen data from the same source, but it cannot detect failures caused by population, site or time differences. External validation on independent data remains uncommon in mental health ML and is one of the most frequently cited gaps [9], [13]. Calibration, meaning the agreement between predicted probabilities and observed outcome rates, is seldom reported even though miscalibrated risks can mislead clinical decisions [36]. Decision curve analysis links predictions to net clinical benefit across threshold choices [37]. Sample size should be planned with prediction-model criteria rather than rules of thumb [38]. The TRIPOD+AI statement now gives consensus guidance for transparent reporting of these elements [39].

7. TRANSLATIONAL CHALLENGES

Fig. 6 maps the principal risks identified in the literature onto the stages of the AI lifecycle. Table 4 pairs them with recommended safeguards.



Upper row: AI lifecycle stage | Lower row: principal risk identified in the literature

© 2026 IJAESCML | All Rights Reserved

International Journal Of Advances In
ENGINEERING, SCIENCE, COMMERCE, MANAGEMENT AND LAW
(IJAESCML)
ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026
www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

Fig. 6. Principal translational risks identified in the literature, mapped onto the AI lifecycle stages from data acquisition to post-deployment monitoring.

7.1 Explainability

Clinicians need to understand why a model flags a person before they can weigh its output. Post-hoc attribution methods such as LIME [40] and SHAP [41], [42] show which features contributed to a prediction. Multidisciplinary analyses argue that explainability supports informed consent, trust and accountability [43]. Others caution that post-hoc explanations can be unfaithful to the model and should not be treated as proof that an individual decision is correct [44], and they recommend inherently interpretable models for high-stakes decisions where these perform comparably [45].

7.2 Fairness and Bias

Models trained on unrepresentative data can perform worse for under-represented groups. A widely used health-management algorithm systematically underestimated the needs of Black patients because it used healthcare cost as a proxy for illness [46]. Mental health data carry comparable risks from cultural differences in symptom expression, language and help-seeking. Fairness therefore needs explicit definitions [47], subgroup evaluation, and design choices aimed at equity [48].

7.3 Dataset Shift, Privacy and Governance

Performance can degrade when deployment conditions differ from development conditions, for example with new devices, new populations or changes in clinical practice. Clinicians need to recognise and report such dataset shift [49]. Mental health information is highly sensitive, so consent, data minimisation, de-identification, encryption and access control are essential. Before routine use, AI tools require staged clinical evaluation under reporting frameworks such as DECIDE-AI [50] and CONSORT-AI [51], along with clear lines of professional accountability.

Table 4. Principal ethical and technical risks with corresponding safeguards

Risk	Manifestation	Recommended safeguard
Label circularity	Model learns the screening questionnaire rather than the disorder	Structured clinical interview as reference standard [16]
Overfitting and leakage	Inflated internal performance	Nested cross-validation; preprocessing inside folds; external test set
Miscalibration	Over- or under-stated risk	Report calibration plots; recalibrate at new sites [36]
Opacity	Unexplained flags erode trust	Interpretable baselines; faithful explanations with caveats [44], [45]
Bias	Unequal error rates across groups	Subgroup metrics; representative sampling [46], [48]
Dataset shift	Silent performance decline	Monitoring, drift detection, scheduled revalidation [49]
Privacy breach	Exposure of sensitive data	Consent, encryption, role-based access, audit trails
Automation bias	Over-reliance on model output	Clinician-in-the-loop; model facts labels [52]

8. RESEARCH GAPS AND FUTURE DIRECTIONS

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823

ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026

www.ijaescml.com | internationaljournal1234@gmail.com | Peer Reviewed & Refereed Journal

The review identifies six priorities for moving AI-assisted diagnosis of depression and anxiety from promising prototypes to clinically useful tools:

- **Reference standards:** use structured diagnostic interviews rather than questionnaire cut-offs as outcome labels, and report the reference standard explicitly.
- **External and temporal validation:** test models on independent sites, later time periods, other languages and other cultural contexts, including under-represented populations in South Asia and other low- and middle-income settings [3].
- **Calibration and utility:** report calibration and decision-analytic measures alongside discrimination [36], [37].
- **Fairness auditing:** evaluate performance across age, gender, language and socioeconomic subgroups as standard practice [48].
- **Human–AI collaboration:** study how clinicians interpret and act on model outputs in prospective, staged evaluations [50].
- **Transparent reporting:** adopt TRIPOD+AI and share code and model documentation where privacy permits [39].

9. CONCLUSION

AI methods can extract clinically relevant signals of depression and anxiety from questionnaires, health records, language, speech and behaviour. Classical models remain strong on structured data, and deep and transformer-based models lead on unstructured data. The evidence base is still constrained by small, homogeneous samples, circular labels, scarce external validation and little attention to calibration and fairness. The most defensible role for AI at present is as clinician-supervised decision support that widens access to screening and focuses professional attention, while diagnosis remains a professional act informed by context, history and dialogue. Realising this role will depend on rigorous validation, explainable and calibrated outputs, fairness auditing and transparent governance.

© 2026 IJAESCML | All Rights Reserved

doi.org/10.5281/zenodo.23167823

REFERENCES

[1] GBD 2019 Mental Disorders Collaborators, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," Lancet Psychiatry, vol. 9, no. 2, pp. 137–150, 2022, doi: 10.1016/S2215-0366(21)00395-3.

[2] COVID-19 Mental Disorders Collaborators, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," Lancet, vol. 398, no. 10312, pp. 1700–1712, 2021, doi: 10.1016/S0140-6736(21)02143-7.

[3] V. Patel et al., "The Lancet Commission on global mental health and sustainable development," Lancet, vol. 392, no. 10157, pp. 1553–1598, 2018, doi: 10.1016/S0140-6736(18)31612-X.

[4] R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters, "Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication," Arch. Gen. Psychiatry, vol. 62, no. 6, pp. 617–627, 2005, doi: 10.1001/archpsyc.62.6.617.

[5] N. H. Kalin, "The critical relationship between anxiety and depression," Am. J. Psychiatry, vol. 177, no. 5, pp. 365–367, 2020, doi: 10.1176/appi.ajp.2020.20030305.

[6] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: A meta-analysis," Lancet, vol. 374, no. 9690, pp. 609–619, 2009, doi: 10.1016/S0140-6736(09)60879-5.

[7] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nat. Med., vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[8] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," Psychol. Med., vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.

[9] Y. Park, S. Park, and M. Lee, "Effectiveness of artificial intelligence in detecting and managing depressive disorders: Systematic review," J. Affect. Disord., vol. 361, pp. 445–456, 2024, doi: 10.1016/j.jad.2024.06.035.

[10] F. Zafar et al., "The role of artificial intelligence in identifying depression and anxiety: A comprehensive literature review," Cureus, vol. 16, no. 3, Art. no. e56472, 2024, doi: 10.7759/cureus.56472.

[11] N. Obradovich et al., "Opportunities and risks of large language models in psychiatry," NPP—Digit. Psychiatry Neurosci., vol. 2, no. 1, Art. no. 8, 2024, doi: 10.1038/s44277-024-00010-z.

[12] A. Pigoni et al., "Machine learning and the prediction of suicide in psychiatric populations: A systematic review," Transl. Psychiatry, vol. 14, no. 1, Art. no. 140, 2024, doi: 10.1038/s41398-024-02852-9.

[13] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," Annu. Rev. Clin. Psychol., vol. 14, pp. 91–118, 2018, doi: 10.1146/annurev-clinpsy-032816-045037.

[14] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Med., vol. 17, Art. no. 195, 2019, doi: 10.1186/s12916-019-1426-2.

[15] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[16] B. Levis, A. Benedetti, and B. D. Thombs, "Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis," BMJ, vol. 365, Art. no. l1476, 2019, doi: 10.1136/bmj.l1476.

[17] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," Arch. Intern. Med., vol. 166, no. 10, pp. 1092–1097, 2006, doi: 10.1001/archinte.166.10.1092.

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

[18] F. Plummer, L. Manea, D. Trepel, and D. McMillan, "Screening for anxiety disorders with the GAD-7 and GAD-2: A systematic review and diagnostic metaanalysis," Gen. Hosp. Psychiatry, vol. 39, pp. 24–31, 2016, doi: 10.1016/j.genhosppsych.2015.11.005.
[19] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," Curr. Opin. Behav. Sci., vol. 18, pp. 43–49, 2017, doi: 10.1016/j.cobeha.2017.07.005.
[20] J. C. Eichstaedt et al., "Facebook language predicts depression in medical records," Proc. Natl. Acad. Sci. USA, vol. 115, no. 44, pp. 11203–11208, 2018, doi: 10.1073/pnas.1802331115.
[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT), 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[22] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," Speech Commun., vol. 71, pp. 10–49, 2015, doi: 10.1016/j.specom.2015.03.004.
[23] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," Laryngoscope Investig. Otolaryngol., vol. 5, no. 1, pp. 96–116, 2020, doi: 10.1002/lio2.354.
[24] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," Annu. Rev. Clin. Psychol., vol. 13, pp. 23–47, 2017, doi: 10.1146/annurev-clinpsy-032816-044949.
[25] S. Saeb et al., "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," J. Med. Internet Res., vol. 17, no. 7, Art. no. e175, 2015, doi: 10.2196/jmir.4273.
[26] J. Torous et al., "The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality," World Psychiatry, vol. 20, no. 3, pp. 318–335, 2021, doi: 10.1002/wps.20883.
[27] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.
[28] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[29] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
[31] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," Nat. Methods, vol. 15, no. 4, pp. 233–234, 2018, doi: 10.1038/nmeth.4642.
[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
[35] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

ISSN (Online)
3048-8427
Vol. 3 | Issue 3
2026

[36] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," BMC Med., vol. 17, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.
[37] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," Med. Decis. Making, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.
[38] R. D. Riley et al., "Calculating the sample size required for developing a clinical prediction model," BMJ, vol. 368, Art. no. m441, 2020, doi: 10.1136/bmj.m441.
[39] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.
[40] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[41] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.
[42] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
[43] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," BMC Med. Inform. Decis. Mak., vol. 20, Art. no. 310, 2020, doi: 10.1186/s12911-020-01332-6.
[44] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit. Health, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.
[45] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
[46] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
[47] S. Verma and J. Rubin, "Fairness definitions explained," in Proc. Int. Workshop Softw. Fairness (FairWare), 2018, pp. 1–7, doi: 10.1145/3194770.3194776.
[48] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," Ann. Intern. Med., vol. 169, no. 12, pp. 866–872, 2018, doi: 10.7326/M18-1990.
[49] S. G. Finlayson et al., "The clinician and dataset shift in artificial intelligence," N. Engl. J. Med., vol. 385, no. 3, pp. 283–286, 2021, doi: 10.1056/NEJMc2104626.
[50] B. Vasey et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," Nat. Med., vol. 28, no. 5, pp. 924–933, 2022, doi: 10.1038/s41591-022-01772-9.
[51] X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension," Nat. Med., vol. 26, no. 9, pp. 1364–1374, 2020, doi: 10.1038/s41591-020-1034-x.
[52] M. P. Sendak, M. Gao, N. Brajer, and S. Balu, "Presenting machine learning model information to clinical end users with model facts labels," npj Digit. Med., vol. 3, Art. no. 41, 2020, doi: 10.1038/s41746-020-0253-3.

© 2026 IJAESCML | All Rights Reserved
www.ijaescml.com

Vol. 3, Issue 3, July-Sseptember 2026
doi.org/10.5281/zenodo.23167823