

A Clinician-in-the-Loop, Explainable and Fairness-Aware Framework for AI-Assisted Screening of Depression and Anxiety: Design and Validation Protocol

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

Artificial intelligence (AI) models for detecting depression and anxiety often report high accuracy in development data but rarely reach routine care, largely because of weak validation, uncalibrated outputs, opaque reasoning, unexamined bias and unclear clinical responsibility. This paper proposes a six-layer framework for AI-assisted screening of depression and anxiety that treats the model as decision support under professional oversight rather than as a diagnostic authority. The layers cover (1) ethical, consent-based multimodal data acquisition anchored on the PHQ-9, GAD-7 and a structured clinical interview; (2) quality control and preprocessing; (3) feature representation for questionnaire, clinical, language, speech and wearable data; (4) modelling with a logistic regression baseline and candidate support vector machine, random forest, gradient boosting and neural network models under nested cross-validation, calibration and external validation; (5) explanation and subgroup fairness auditing; and (6) a clinician-in-the-loop decision pathway with feedback for monitoring. A cross-cutting governance layer addresses privacy, security and accountability. A complete evaluation protocol is specified, including sample size planning, metric definitions, threshold selection and reporting under TRIPOD+AI. An analytic demonstration using published instrument accuracies shows why prevalence-aware interpretation is essential. At 5% prevalence, a screen with 0.88 sensitivity and 0.88 specificity yields a positive predictive value of only 27.8%, rising to 64.7% at 20% prevalence, while negative predictive value stays above 96%. The framework gives researchers and clinicians a reproducible template for developing, validating and deploying AI screening tools responsibly.

Keywords— *AI-assisted screening; depression; anxiety; clinical decision support; human-in-the-loop; explainable AI; algorithmic fairness; model calibration; validation protocol; TRIPOD+AI*

1. INTRODUCTION

1.1 Motivation

Depressive and anxiety disorders affect an estimated 280 million and 301 million people worldwide, respectively [1], and their prevalence rose sharply during the COVID-19 pandemic [2]. Many affected individuals are never identified. In primary care, clinicians recognise only about half of patients with depression [3], and anxiety is frequently overshadowed by co-occurring mood or somatic complaints [4]. Brief instruments such as the PHQ-9 and GAD-7 perform well [5], [6], but they depend on accurate self-report and on staff with the time to administer and interpret them.

Machine learning (ML) can combine many weak indicators, such as questionnaire items, clinical records, language, speech and daily behaviour, into a single risk estimate [7], [8]. Recent reviews document rapid growth in such models [9], [10] and the arrival of large language

models (LLMs) in psychiatry [11]. They also document recurring weaknesses: small single-site samples, questionnaire scores used as ground truth, scarce external validation, rarely reported calibration, and almost no fairness evaluation [12], [13]. These weaknesses explain why promising prototypes seldom become trusted clinical tools.

1.2 Contributions

This paper derives from the doctoral thesis *AI Assisted Diagnosis of Depression and Anxiety*. It converts that thesis's methodological design into a reusable framework, with four contributions:

- A six-layer architecture for AI-assisted screening that places clinical judgement at the decision point and treats governance as a cross-cutting concern (Section 3).

- A complete validation and evaluation protocol covering sample size, nested and external validation, calibration, threshold selection and reporting (Section 4).
- An explicit explanation, fairness-audit and clinician-in-the-loop pathway (Sections 5 and 6).
- An analytic demonstration, based on published instrument accuracies, of how prevalence and threshold choice shape the clinical meaning of a model output (Section 7).

[15], speech [16], [17] and smartphone or wearable sensing [18], [19]. Classical algorithms such as support vector machines [20], random forests [21] and gradient boosting [22] remain competitive on tabular data, while transformer language models [23] lead on text. Methodological guidance has matured in parallel: TRIPOD+AI for reporting [24], principled sample-size calculation [25], calibration assessment [26], and staged clinical evaluation through DECIDE-AI [27]. Table 1 contrasts common practice with what the proposed framework requires.

2. RELATED WORK AND GAP ANALYSIS

Studies of AI-assisted detection have drawn on electronic health records, social media language [14],

Table 1. Gap analysis: common practice in the literature versus requirements of the proposed framework

Dimension	Common practice	Framework requirement
Reference standard	Questionnaire cut-off used as label	Structured clinical interview; questionnaire used as a feature [28]
Sample size	Convenience sample	Prediction-model sample-size criteria [25]
Validation	Single random split	Nested cross-validation plus temporal and external test sets
Performance	Accuracy, AUC only	Sensitivity, specificity, PPV, NPV, PR-AUC, calibration, net benefit [29], [30]
Explainability	Absent or ad hoc	Interpretable baseline plus SHAP-style attribution with stated caveats [31], [32]
Fairness	Not assessed	Subgroup metrics by age, gender, language and data source [33]
Clinical role	Implicitly autonomous	Clinician-in-the-loop with logged overrides [27]
Reporting	Variable	TRIPOD+AI checklist [24]

3. PROPOSED FRAMEWORK

3.1 Architecture Overview

Fig. 1 presents the framework as six sequential layers and a vertical governance layer. Each layer produces documented outputs that feed the next, so that every step from data collection to clinical decision can be audited.

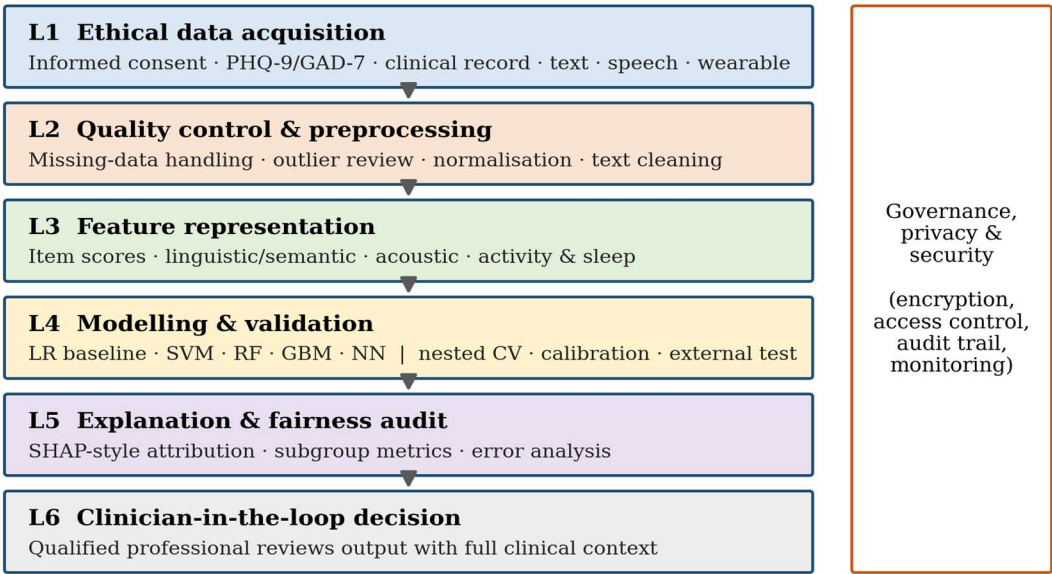


Fig. 1. Proposed six-layer framework for AI-assisted screening of depression and anxiety. The governance, privacy and security layer applies across all stages.

3.2 Layer 1: Ethical Data Acquisition

Adult participants are recruited under institutional ethics approval with written informed consent that explains the purpose of data collection, which data are gathered, how they are stored, and the right to withdraw. The core dataset comprises sociodemographic variables, PHQ-9 [5] and GAD-7 [6] item responses, and relevant clinical history. Optional modalities, collected only with separate consent, are free-text responses to open prompts, short speech recordings, and wearable-derived activity and sleep summaries. The outcome label is a diagnosis established by structured clinical interview and blinded to model inputs. Using the questionnaire as a predictor rather than as the label avoids the circularity that affects much of the literature [28]. Individuals who show acute risk at any point are referred immediately under a pre-specified safety protocol.

3.3 Layer 2: Quality Control and Preprocessing

Data are checked for completeness, range violations and duplicates. Missing values are summarised by variable and subgroup, and they are handled by multiple

imputation fitted within training folds only. Outliers are reviewed for plausibility rather than removed automatically. Continuous features are standardised with parameters learned on training data. Text is de-identified, normalised and tokenised. Audio is resampled, silence-trimmed and screened for signal quality. Every transformation is fitted inside the cross-validation loop to prevent information leakage.

3.4 Layer 3: Feature Representation

Table 2 lists the feature groups for each modality. Questionnaire features include item scores, total scores and response patterns. Language features include lexical, affective and pronoun-use indices [14] as well as contextual embeddings from pretrained transformers [23]. Speech features include prosodic, spectral and voice-quality descriptors associated with psychomotor change [16]. Wearable features summarise mobility, activity regularity and sleep [18], [19]. Feature selection uses penalised regression and stability selection within training folds.

Table 2. Feature groups and preprocessing by data modality

Modality	Feature group	Key preprocessing
Questionnaire	PHQ-9 and GAD-7 items, totals, response latency	Range checks; imputation inside folds
Clinical / sociodemographic	Age, gender, history, medication, comorbidity	Coding harmonisation; one-hot encoding
Language	Lexical and affective indices; transformer embeddings	De-identification; normalisation; tokenisation
Speech	Pitch, energy, pause, speech rate, spectral features	Resampling; noise and quality screening
Wearable	Step count, activity regularity, sleep duration and timing	Wear-time validation; daily aggregation

3.5 Layer 4: Candidate Models

Regularised logistic regression serves as the transparent baseline. Candidate models are a support vector machine [20], a random forest [21], gradient boosting [22], and a feed-forward or recurrent neural network [34], [35] for sequential modalities. Multimodal inputs are combined by late fusion of modality-specific probabilities by default, which keeps each modality auditable. A more complex model replaces the baseline only if it improves discrimination and net benefit on external data without degrading calibration or subgroup fairness [36], [37]. Class imbalance is handled by class weighting, or by SMOTE applied within training folds only [38].

4. VALIDATION AND EVALUATION PROTOCOL

4.1 Sample Size Planning

The required development sample is calculated using the criteria of Riley *et al.* [25]: precise estimation of the

overall outcome proportion, a target expected shrinkage of at least 0.9 to limit overfitting, and a small optimism in apparent R^2 . These criteria depend on the anticipated outcome prevalence and on the number of candidate predictor parameters, so they discourage feature-heavy models on small samples.

4.2 Validation Design

As shown in Fig. 2, performance is estimated by stratified nested cross-validation. The outer five folds give an approximately unbiased estimate of generalisation. The inner folds tune hyperparameters, select features and fit imputation models. The final model is then assessed on a temporal hold-out drawn from a later recruitment period and, where collaboration allows, on an external sample from an independent site or language group. External validation is treated as essential rather than optional, because internal validation cannot reveal dataset shift [39].

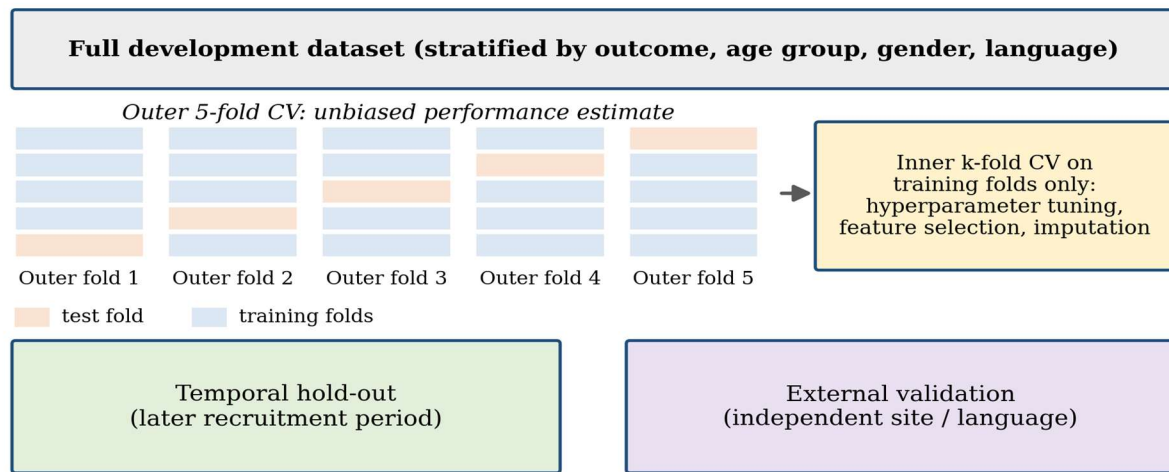


Fig. 2. Validation design. Stratified nested cross-validation separates tuning from performance estimation, followed by temporal and external validation.

4.3 Performance Metrics

Let TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, and let p be the prevalence. The primary metrics are defined in Table 3. Predictive values follow from Bayes' theorem:

$$PPV = \frac{(Se \cdot p)}{[Se \cdot p + (1 - Sp) \cdot (1 - p)]} \quad (1)$$

$$NPV = \frac{[Sp \cdot (1 - p)]}{[Sp \cdot (1 - p) + (1 - Se) \cdot p]} \quad (2)$$

Table 3. Evaluation metrics and their role in the protocol

Metric	Definition	Role in the protocol
Sensitivity (Se)	TP / (TP + FN)	Primary for screening; missed cases are costly
Specificity (Sp)	TN / (TN + FP)	Controls unnecessary referrals
PPV / NPV	Equations (1) and (2)	Reported at the target-setting prevalence
F1 score	$2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$	Summary under imbalance
ROC-AUC	Area under the ROC curve	Threshold-free discrimination
PR-AUC	Area under the precision-recall curve	Informative under imbalance [29]
Calibration	Intercept, slope, calibration plot	Trustworthiness of predicted risk [26]
Net benefit	Decision curve analysis	Clinical utility across thresholds [30]

4.4 Calibration and Threshold Selection

Probabilities are calibrated with Platt scaling or isotonic regression fitted on inner folds. Calibration intercept, slope and plots are reported for every validation set [26]. Operating thresholds are selected on development data and then frozen before external testing. For a first-line screen, the threshold is chosen to achieve a pre-specified sensitivity, for example 0.90, so that few cases are missed. The Youden index is reported for comparison, and decision curves show the range of thresholds over which the model offers net benefit [30].

5. EXPLAINABILITY AND CLINICIAN-IN-THE-LOOP DECISION MAKING

5.1 Explanation Layer

Each prediction is accompanied by a global and a local explanation. Globally, logistic regression coefficients and model-agnostic permutation importance summarise which features drive the model. Locally,

SHAP values [31], [40] show how each feature moved an individual's risk up or down. LIME [41] is used as a sensitivity check. Explanations are presented as descriptions of model behaviour, not as causal or clinical reasons, because post-hoc explanations can be unfaithful [32]. A model facts label [42] summarises intended use, training population, performance, subgroup results and known limitations for clinical users [43].

5.2 Clinician-in-the-Loop Pathway

Fig. 3 shows the decision pathway. The model output, a calibrated probability with its explanation, goes to a qualified professional. The professional integrates it with history, observation and interview and assigns one of three actions: routine follow-up with psychoeducation, structured diagnostic interview, or urgent specialist assessment. The clinician may override the model at any point, and overrides, reasons and eventual outcomes are logged. This log guards against automation bias and supplies data for monitoring and recalibration.

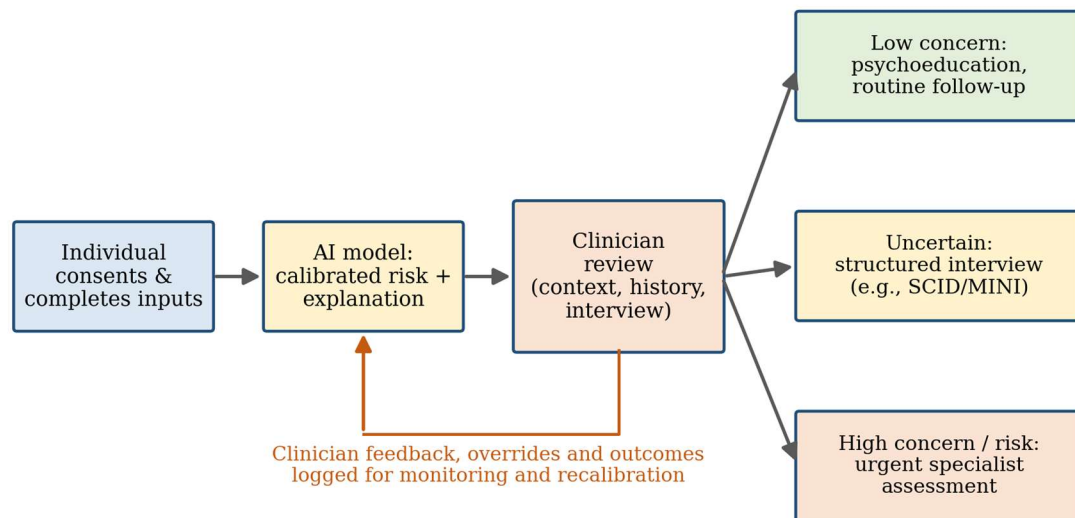


Fig. 3. Clinician-in-the-loop decision pathway. The AI output informs, but does not replace, professional assessment, and clinician feedback is logged to support monitoring and recalibration.

6. FAIRNESS AUDITING AND GOVERNANCE

6.1 Subgroup Fairness Audit

Algorithms can encode structural inequities through proxy labels or unrepresentative data [44]. The framework therefore requires a pre-specified fairness audit (Table 4). Sensitivity, specificity, PPV and calibration are computed for subgroups defined by age

band, gender, language of assessment and data source. Differences are reported with confidence intervals and judged against fairness definitions agreed in advance [45]. Where material disparities appear, mitigation through reweighting, subgroup-specific thresholds or additional data collection is carried out and re-evaluated [33].

Table 4. Pre-specified fairness audit plan

Subgroup axis	Metrics compared	Action if disparity is detected
Age band	Sensitivity, specificity, calibration slope	Reweight; collect more data from under-represented ages
Gender	Equal opportunity (sensitivity gap), PPV	Review features; consider group-specific thresholds
Language of assessment	Sensitivity, PR-AUC, calibration	Validate translated instruments; language-specific language models
Data source / site	All primary metrics	Recalibrate per site; examine dataset shift [39]

6.2 Privacy, Security and Accountability

Data are pseudonymised at collection, encrypted at rest and in transit, and held on access-controlled institutional servers with audit trails. Only the minimum data required are retained, and raw audio and text are deleted after feature extraction where the protocol allows. Model versions, training data snapshots and evaluation reports are archived. Responsibility for clinical decisions remains with the treating professional, while model developers are responsible for documented performance and for timely disclosure of known limitations [12].

7. ANALYTIC DEMONSTRATION

7.1 Effect of Prevalence on Predictive Values

To show why prevalence-aware interpretation is required, equations (1) and (2) were evaluated using the published operating characteristics of the PHQ-9 (Se = Sp = 0.88) [5] and the GAD-7 (Se = 0.89, Sp = 0.82) [6]. The same accuracy applies to any AI model with equivalent sensitivity and specificity. As Fig. 4 and Table 5 show, at 5% prevalence, typical of general-population screening, only 27.8% of positive PHQ-9 screens are true cases. At 20% prevalence, typical of primary-care attenders with psychological complaints, the figure rises to 64.7%. Negative predictive value stays above 96% in both settings. A positive AI flag in a low-prevalence setting therefore justifies further assessment, not a diagnosis, while a negative flag is reassuring.

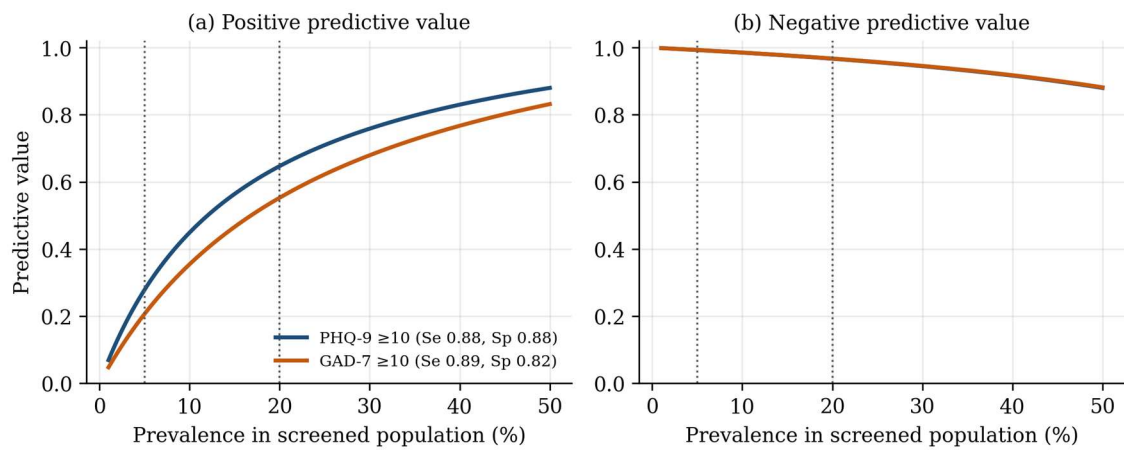


Fig. 4. Positive (a) and negative (b) predictive values as a function of prevalence, computed from equations (1) and (2) using published PHQ-9 [5] and GAD-7 [6] accuracy. Dotted lines mark 5% and 20% prevalence.

Table 5. Predictive values at two prevalence levels (computed from equations (1) and (2))

Instrument / operating point	PPV at 5%	PPV at 20%	NPV at 5%	NPV at 20%
PHQ-9 ≥ 10 (Se 0.88, Sp 0.88)	27.8%	64.7%	99.3%	96.7%
GAD-7 ≥ 10 (Se 0.89, Sp 0.82)	20.6%	55.3%	99.3%	96.8%

7.2 Threshold Trade-off

Fig. 5 illustrates threshold selection for a hypothetical model with ROC-AUC 0.80 under a binormal assumption. This is an analytic illustration, not an empirical result. Setting the threshold to achieve a sensitivity of 0.90 lowers specificity to about 0.46. The Youden-optimal point balances the two at about 0.72

each. For a first-stage screen, where missed cases carry high cost and a positive result leads only to clinical review, the high-sensitivity point is preferable. The framework therefore fixes thresholds in advance according to the intended clinical role instead of reporting a single default cut-off.

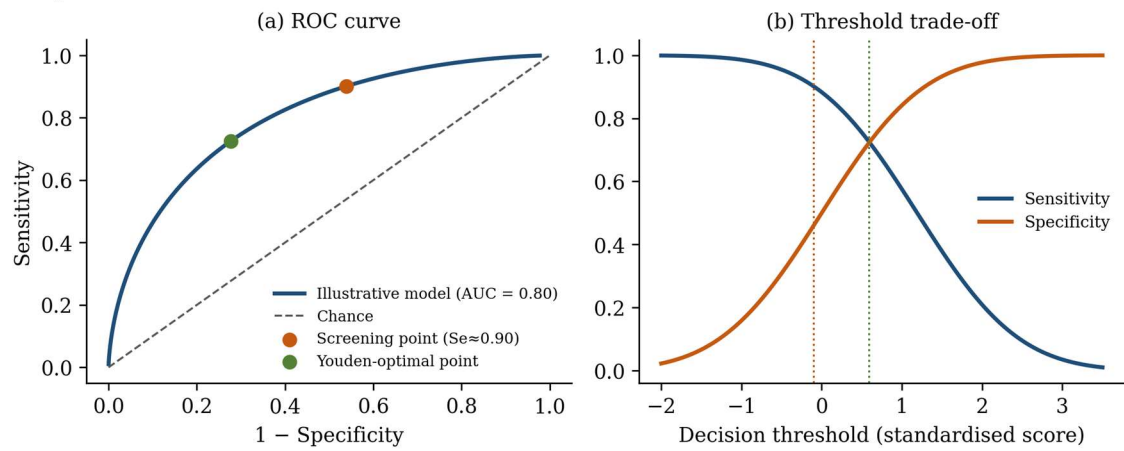


Fig. 5. Illustrative threshold trade-off for a hypothetical model (binormal, AUC = 0.80). (a) ROC curve with a high-sensitivity screening point and the Youden-optimal point. (b) Sensitivity and specificity as functions of the decision threshold.

8. IMPLEMENTATION AND LIFECYCLE MONITORING

Deployment follows the staged lifecycle in Fig. 6. After external validation, the tool enters a supervised early clinical evaluation reported under DECIDE-AI [27], in which usability, workflow fit, clinician trust and safety are assessed before any comparative trial under

CONSORT-AI [46]. In routine use, discrimination, calibration, subgroup metrics, override rates and input-data distributions are monitored on a schedule. Pre-defined triggers prompt recalibration or retraining when drift is detected [39]. Throughout, development and validation are reported with the TRIPOD+AI checklist [24].

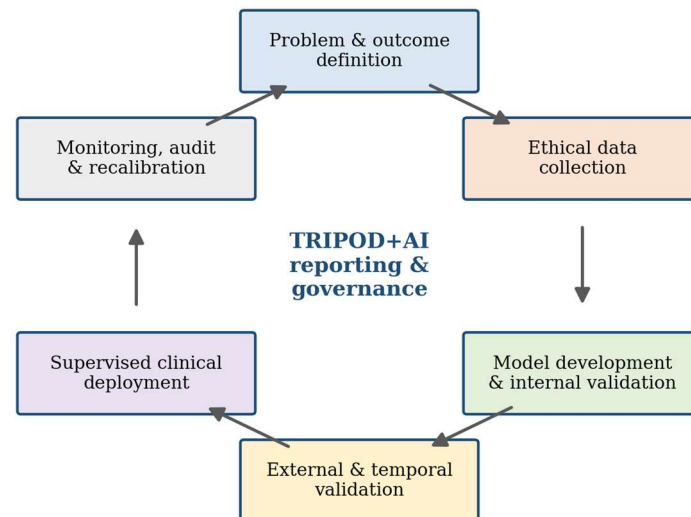


Fig. 6. Lifecycle of the AI screening tool, from problem definition to monitoring and recalibration, with TRIPOD+AI reporting and governance at its centre.

9. DISCUSSION

9.1 Implications

The framework shifts attention from maximising a single accuracy figure to building a tool whose outputs are well calibrated, explainable, fair across groups and embedded in accountable clinical decisions. This is consistent with evidence that a validated questionnaire used well already achieves high accuracy [28], [47], so the added value of AI lies in scale, consistency, multimodal integration and continuous monitoring rather than in replacing clinical assessment [48], [49]. For settings with limited specialist capacity, such as much of South and North-East India, a calibrated, clinician-supervised screen could help direct scarce professional time to those most in need [50].

9.2 Limitations

The framework is a design and protocol, and it has not yet been validated on a participant dataset. The analytic demonstration relies on published instrument characteristics and an idealised binormal model, and real models will show different curves. Multimodal collection raises burden and privacy concerns that may limit participation. LLM-based components, while promising, introduce further risks of hallucination and data exposure that need dedicated evaluation [11]. Finally, fairness auditing depends on adequate subgroup sample sizes, which may require multi-site collaboration.

10. CONCLUSION

This paper has presented a six-layer, clinician-in-the-loop framework for AI-assisted screening of depression and anxiety, together with a complete validation, explanation, fairness and governance protocol. The

central principle is that AI output is an additional, calibrated and explained source of information that supports, and never replaces, professional judgement. The analytic demonstration shows that even accurate screens yield modest positive predictive values at low prevalence, which reinforces the need for clinical confirmation. Applying the framework to an ethically collected multimodal dataset, with external validation and prospective clinical evaluation, is the next step toward safe and equitable AI support in mental health care.

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](https://doi.org/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](https://doi.org/10.1016/S0140-6736(21)02143-7).
- [3] 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](https://doi.org/10.1016/S0140-6736(09)60879-5).
- [4] 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](https://doi.org/10.1176/appi.ajp.2020.20030305).
- [5] 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](https://doi.org/10.1046/j.1525-1497.2001.016009606.x).
- [6] 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](https://doi.org/10.1001/archinte.166.10.1092).
- [7] 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](https://doi.org/10.1017/S0033291719000151).
- [8] 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](https://doi.org/10.1146/annurev-clinpsy-032816-045037).
- [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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1038/s44277-024-00010-z).
- [12] 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](https://doi.org/10.1186/s12916-019-1426-2).
- [13] A. M. Chekroud et al., "The promise of machine learning in predicting treatment outcomes in psychiatry," *World Psychiatry*, vol. 20, no. 2, pp. 154–170, 2021, doi: [10.1002/wps.20882](https://doi.org/10.1002/wps.20882).
- [14] 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](https://doi.org/10.1016/j.cobeha.2017.07.005).
- [15] 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](https://doi.org/10.1073/pnas.1802331115).
- [16] 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](https://doi.org/10.1016/j.specom.2015.03.004).
- [17] 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](https://doi.org/10.1002/lio2.354).
- [18] 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](https://doi.org/10.1146/annurev-clinpsy-032816-044949).
- [19] 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](https://doi.org/10.2196/jmir.4273).
- [20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: [10.1007/BF00994018](https://doi.org/10.1007/BF00994018).
- [21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- [22] 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](https://doi.org/10.1214/aos/1013203451).
- [23] 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](https://doi.org/10.18653/v1/N19-1423).
- [24] 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](https://doi.org/10.1136/bmj-2023-078378).
- [25] 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](https://doi.org/10.1136/bmj.m441).
- [26] 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](https://doi.org/10.1186/s12916-019-1466-7).
- [27] 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](https://doi.org/10.1038/s41591-022-01772-9).
- [28] 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. 11476, 2019, doi: [10.1136/bmj.11476](https://doi.org/10.1136/bmj.11476).
- [29] 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](https://doi.org/10.1371/journal.pone.0118432).
- [30] 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](https://doi.org/10.1177/0272989X06295361).
- [31] 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](https://doi.org/10.48550/arXiv.1705.07874).
- [32] 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](https://doi.org/10.1016/S2589-7500(21)00208-9).
- [33] 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](https://doi.org/10.7326/M18-1990).
- [34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: [10.1038/nature14539](https://doi.org/10.1038/nature14539).
- [35] 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](https://doi.org/10.1162/neco.1997.9.8.1735).
- [36] 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](https://doi.org/10.1038/s42256-019-0048-x).
- [37] 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](https://doi.org/10.1038/nmeth.4642).
- [38] 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](https://doi.org/10.1613/jair.953).
- [39] 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](https://doi.org/10.1056/NEJMc2104626).
- [40] 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](https://doi.org/10.1038/s42256-019-0138-9).
- [41] 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](https://doi.org/10.1145/2939672.2939778).
- [42] 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](https://doi.org/10.1038/s41746-020-0253-3).
- [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](https://doi.org/10.1186/s12911-020-01332-6).

- [44] 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](https://doi.org/10.1126/science.aax2342).
- [45] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. Int. Workshop Softw. Fairness (FairWare)*, 2018, pp. 1–7, doi: [10.1145/3194770.3194776](https://doi.org/10.1145/3194770.3194776).
- [46] 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](https://doi.org/10.1038/s41591-020-1034-x).
- [47] 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](https://doi.org/10.1016/j.genhosppsych.2015.11.005).
- [48] 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](https://doi.org/10.1038/s41591-018-0300-7).
- [49] 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](https://doi.org/10.1002/wps.20883).
- [50] 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](https://doi.org/10.1016/S0140-6736(18)31612-X).