



Explainable AI in Mental Health Decision Systems: A Structured Review of Translational Maturity

Pramesti Dewi¹, Annastasya Nabila Elsa Wulandari², Purwono³, Indah Trivilia⁴

¹ *Department of Nursing, Universitas Harapan Bangsa, Purwokerto, Indonesia*

^{2,3} *Department of Informatics, Universitas Harapan Bangsa, Purwokerto, Indonesia*

⁴ *Department of Information System, Universitas Harapan Bangsa, Purwokerto, Indonesia*

ARTICLE INFO

Article history:

Received February 19, 2026

Revised August 27, 2026

Published August 29, 2026

Keywords:

Explainable AI;
External Validation;
Mental Health;
Clinical Translation;
Human-Centered Evaluation

ABSTRACT

The application of Artificial Intelligence (AI) in mental health is expanding, but explainability alone does not establish readiness for clinical use. This structured review synthesizes empirical studies of Explainable Artificial Intelligence (XAI) in mental health decision systems through three review-specific dimensions: explainability depth, validation strength, and clinical integration. The search followed PRISMA 2020 reporting principles across IEEE Xplore, PubMed, ScienceDirect, Scopus, and SpringerLink for English-language journal articles published from 2020 to 2026. Nine studies met the predefined eligibility criteria. Within this narrowly selected evidence set, post-hoc methods, especially SHAP, predominated, whereas intrinsic and counterfactual approaches were less frequent. Most studies relied on internal validation, only one reported independent external validation, and prospective workflow deployment was not identified. Human-centered evaluation of explanations was also uncommon. To structure these observations, we operationalize a three-axis ordinal profile (E1–E3, V1–V3, C1–C3) covering explainability depth, validation strength, and clinical integration. This profile is intended as a review-specific descriptive instrument rather than a validated clinical-readiness score. Its contribution is to integrate these three dimensions at the individual-study level, complementing rather than replacing established reporting, implementation, and clinical AI evaluation frameworks. The findings suggest that the selected literature is more mature in computational transparency than in external generalization and workflow integration. Future studies should combine explanation fidelity and stability assessment, external validation, prospective human-in-the-loop evaluation, and post-deployment monitoring.

This work is licensed under a [Creative Commons Attribution-Share Alike 4.0](https://creativecommons.org/licenses/by-sa/4.0/)



Corresponding Author:

Purwono, Universitas Harapan Bangsa, Indonesia

Email: purwono@uhb.ac.id

1. INTRODUCTION

The use of Artificial Intelligence (AI) in mental health is growing rapidly, particularly in high-risk areas such as diagnosis, risk stratification, and clinical intervention planning [1]. When models function as black-box systems with non-transparent decision-making processes, the clinical implications become even more serious due to limitations in explainability and interpretability, especially in complex architectures such as deep

learning [2]. A number of studies confirm that internal model transparency is an important prerequisite for building trust and ensuring safety in sensitive medical applications [3], [4]. In high-stakes contexts such as suicide, dementia, and depression, misclassification not only affects statistical performance, but also patient safety and the quality of healthcare services. The potential for bias stemming from cultural, gender, and social determinants of health further complicates model interpretation, especially when symptoms overlap between disorders and exhibit high heterogeneity [5], [6]. Unlike many biomedical applications that rely on objective biomarkers, psychiatric diagnoses are contextual and multidimensional, requiring AI systems to not only achieve predictive accuracy but also generate explanations that are verifiable and meaningful to clinicians [7]. Algorithmic transparency thus serves as a prerequisite for responsible clinical integration, not merely an additional technical attribute [8], [9].

In a broader framework, mental health can be positioned as part of biomedical decision systems, which are data-driven systems that support diagnostic, prognostic, and therapeutic decisions [10]. The complexity of diagnosis, high clinical uncertainty, and ethical sensitivity make mental health a representative domain for testing the readiness of AI translation in the biomedical context in general [11]. With this function, mental health acts as an analytical lens that reveals the challenges of integrating explainability, methodological validation, and clinical accountability that are also relevant to other biomedical decision systems [12]. Although the empirical evidence in this study focuses on mental health applications, the conceptual implications of the developed translation framework have a broader scope for biomedical decision systems as a whole.

Previous reviews have established important conceptual and methodological foundations for XAI in healthcare. Markus et al. [13] organized explainability around terminology, design choices, and evaluation strategies, emphasizing the distinction between interpretable modeling and post-hoc explanation. A broader systematic review by Loh et al. [14] mapped the application of XAI techniques across healthcare domains, while Jung et al. [15] examined the essential properties of explanations and the effectiveness of their evaluation. In mental health specifically, Joyce et al. [4] proposed the Transparency and Interpretability for Understandability (TIFU) framework, which conceptualizes understandability through transparency and interpretability. Importantly, reliability, predictive validity, and generalizability were treated as related but distinct concerns rather than integrated components of the explainability framework. Collectively, these studies provide substantial insight into what constitutes explainability and how explanations may be generated or evaluated, but they do not jointly characterize explainability architecture, methodological validation, and clinical integration at the individual-study level.

A parallel body of literature addresses the translation and implementation of digital and AI-enabled health technologies from perspectives that extend beyond explainability. The NASSS framework examines nonadoption, abandonment, scale-up, spread, and sustainability of health technologies across technological, organizational, adopter, and wider-system domains [16]. AI-specific reporting and evaluation guidance subsequently addressed different stages of clinical evidence generation. CONSORT-AI provides reporting guidance for randomized clinical trials involving AI interventions [17], DECIDE-AI focuses on early-stage clinical evaluation of AI-driven decision-support systems in live clinical settings [18], and TRIPOD+AI provides updated reporting guidance for prediction models developed using regression or machine-learning methods [19]. More recently, FUTURE-AI has articulated lifecycle-oriented principles and best practices for trustworthy and deployable healthcare AI, including fairness, universality, traceability, usability, robustness, and explainability [20]. End-to-end implementation models such as SALIENT further connect model development, data pipelines, human-computer interaction, and clinical workflows across implementation stages [21]. These frameworks provide complementary perspectives on implementation, reporting, validation, and trustworthy deployment, but they were not designed as XAI-specific instruments for jointly profiling explanation characteristics, validation strength, and clinical integration within mental health decision systems.

Against this background, the contribution of the present review is deliberately narrower than proposing another general implementation or clinical-readiness framework. Instead, it introduces a review-specific descriptive profile that integrates three analytically distinct dimensions: explainability depth, methodological validation strength, and clinical integration. These dimensions are examined concurrently for each included study to identify whether advances in computational transparency are accompanied by evidence of generalizability and meaningful clinical embedding. The proposed profile therefore complements rather than replaces established reporting, implementation, and trustworthy-AI frameworks. It is not intended as a validated clinical-readiness score, but as a structured analytical lens for identifying translational asymmetries within the selected mental health XAI literature. Accordingly, this review aims to characterize the application of XAI in mental health decision systems, examine the relationship between explainability approaches and

validation practices, assess the extent of clinical integration, and identify structural barriers that may limit translation into clinical practice.

2. METHODS

2.1. Protocol & Reporting Standard

This study employed a structured review design with systematic literature identification, predefined eligibility criteria, structured data extraction, and methodological quality appraisal. The identification and selection of studies were reported in accordance with the PRISMA 2020 principles [22]. The review was designed to provide an analytical synthesis rather than a quantitative meta-analysis, with particular emphasis on the relationship between explainability, methodological validation, and clinical integration in mental health AI systems.

The publication period was restricted a priori to January 2020 through January 2026. The lower boundary of 2020 was selected to capture the contemporary phase of clinical AI evaluation in which formal AI-specific reporting and evaluation guidance began to emerge, including SPIRIT-AI [23] and CONSORT-AI [24] in 2020 and subsequent guidance for early-stage clinical AI evaluation such as DECIDE-AI [25]. This period also corresponds to the rapid expansion of clinically oriented explainable AI research and provides a focused window for examining current translational practices rather than the historical development of interpretability methods. The databases searched were IEEE Xplore, PubMed, ScienceDirect, Scopus, and SpringerLink to provide complementary coverage of biomedical, clinical, and computational literature.

2.2. Eligibility Criteria

Eligibility criteria were established prior to the screening stage to ensure consistency, transparency, and reproducibility of literature selection and to align the scope of the review with the objectives of evaluating XAI translation in clinically relevant mental health and biomedical decision-making systems. The included studies are at the intersection of the domains of mental health, artificial intelligence applications, and the use of explicitly defined and clinically meaningful explainability mechanisms. XAI is operationally defined as interpretability or model explanation techniques that reveal feature contributions, prediction processes, or decision rationales through post-hoc or intrinsic model approaches with direct relevance to diagnosis, prognosis, risk stratification, or clinical decision support in real-world practice. Purely methodological, simulative, or interpretability-focused research without an operational clinical context is excluded. This strategy prioritizes applications with concrete translational implications to ensure that the review assesses the clinical implementation readiness of XAI rather than merely the technical existence of interpretability techniques, as summarized in Table 1.

Table 1. Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Peer-reviewed empirical journal articles (English, 2020–2026)	Editorials, commentaries, reviews, conference abstracts, non-English
Explicit focus on mental health or psychiatric conditions	General healthcare without mental health focus
AI/ML/DL models with explicit explainability mechanism (intrinsic or post-hoc)	Black-box models without defined explanation method
Supports diagnosis, prognosis, risk prediction, or clinical decision support	Purely technical applications without clinical linkage
Real patient or clinically relevant datasets	Simulated or purely methodological studies

Conference papers and conference abstracts were excluded because the review was specifically designed to evaluate clinically oriented empirical evidence for which sufficiently detailed reporting of participants, predictors, outcomes, analytical procedures, validation strategies, and clinical implementation was required. This restriction was intended to support consistent data extraction and risk-of-bias appraisal across the included evidence base. It should not be interpreted as implying that conference research is methodologically inferior, particularly because technically influential XAI research is frequently disseminated through computer science conferences. Consequently, the exclusion of conference publications represents a deliberate scope restriction and may have omitted relevant methodological innovations that had not yet been reported in full-length peer-reviewed journal articles.

2.3. Search Strategy

The search strategy was designed to identify empirical studies on the intersection of XAI, AI, machine learning, and clinical decision-making applications in mental health. The search was conducted on IEEE Xplore, PubMed, ScienceDirect, Scopus, and SpringerLink using a combination of Boolean terms: ("explainable artificial intelligence" OR "interpretable machine learning" OR "model explanation" OR "post-hoc explanation" OR "intrinsic interpretability" OR SHAP OR LIME OR "counterfactual - explanation") AND ("mental health" OR "psychiatric disorder" OR depression OR anxiety OR suicide OR "cognitive decline" OR dementia OR "severe mental illness") AND ("clinical decision support" OR "risk prediction" OR diagnosis OR prognosis OR screening OR monitoring). The syntax was adjusted to each database's indexing scheme to ensure query compatibility and search sensitivity. The search scope was limited to English-language empirical journal articles from 2020–2026 to represent the latest developments in XAI in clinical decision systems. This strategy focused on applications with an operational translation context, thus excluding technical or conceptual studies without a direct link to clinical practice.

2.4. Study Selection

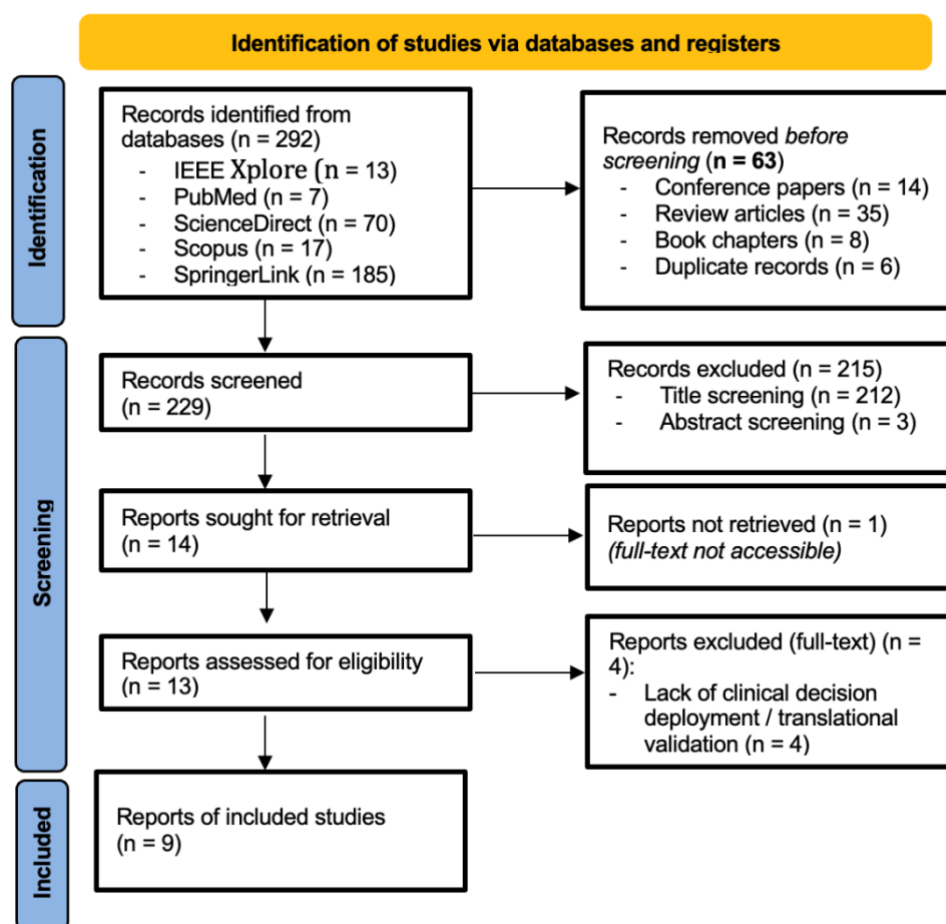


Fig. 1. PRISMA flow diagram

All retrieved records underwent automated deduplication followed by manual verification. Title and abstract screening were then conducted according to the predefined eligibility criteria, with particular attention to mental health relevance, the use of AI or machine-learning models, and the presence of an explicitly defined explainability mechanism. Articles retained after initial screening underwent full-text assessment to determine whether they satisfied the operational definitions of explainability, empirical clinical relevance, and applicability to diagnosis, prognosis, risk stratification, screening, monitoring, or clinical decision support.

Screening, full-text eligibility assessment, data extraction, and subsequent methodological appraisal were conducted by a single reviewer. Duplicate independent screening or coding by a second reviewer was not

performed; therefore, inter-rater agreement statistics were not applicable and were not estimated. To improve procedural consistency, the reviewer used predefined eligibility criteria and a structured extraction template, and full-text inclusion decisions were checked a second time against the eligibility criteria before final inclusion. These procedures improve traceability and internal consistency but cannot eliminate reviewer-dependent selection or interpretation bias. This limitation was therefore considered explicitly when interpreting the findings. Details of the study identification, screening, and inclusion flow are shown in [Figure 1](#).

2.5. Data Extraction

Data were extracted using a predefined structured extraction form. For each included study, bibliographic information, mental health domain, study population, sample size, model architecture, explainability method, validation strategy, external validation status, human-centered evaluation, and clinical implementation stage were recorded. Additional information relevant to methodological appraisal, including study design, predictor definition, outcome assessment, calibration reporting, missing-data handling, class imbalance, and potential overfitting, was extracted when reported.

Because data extraction was undertaken by one reviewer, no inter-rater reliability statistic was calculated. To reduce coding drift, operational definitions for the principal variables were specified before full-text extraction and applied consistently across all included studies. Entries used for the final comparative tables were subsequently rechecked against the corresponding full-text reports. Nevertheless, this procedure constitutes within-reviewer verification rather than independent duplicate extraction and should not be interpreted as equivalent to inter-rater validation.

2.6. Quality Assessment and Operationalization

Risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST), which was developed for studies that develop, validate, or update diagnostic and prognostic prediction models [26]. PROBAST evaluates four domains: participants, predictors, outcome, and analysis. Domain-level signaling questions were used to inform judgments of low, high, or unclear risk of bias in accordance with the PROBAST framework.

The appraisal considered participant selection and representativeness, predictor definition and measurement, outcome determination, analytical procedures, sample size and event adequacy, missing-data handling, overfitting, calibration, and validation strategy when these elements were reported in the primary studies. Particular attention was given to the analysis domain because inadequate validation, small effective sample sizes, inappropriate handling of predictors or outcomes, and overfitting can substantially affect the credibility of prediction-model performance.

PROBAST assessment was conducted by the same reviewer responsible for study selection and data extraction. No independent second appraisal was available and, consequently, inter-rater reliability was not estimated. Domain judgments were rechecked against the corresponding signaling questions before the final study-level classification was assigned. These judgments should therefore be interpreted as structured single-reviewer assessments rather than consensus-based risk-of-bias adjudications.

2.7 Operationalization of Translational Maturity

To structure the comparative synthesis, translational maturity was operationalized across three analytically distinct dimensions: explainability depth (E), validation strength (V), and clinical integration (C). These dimensions were informed by established literature on explainable AI evaluation, prediction-model validation, and clinical AI evaluation [18], [19], [25]. The operational levels themselves were developed specifically for this review and were not adopted from an existing validated scoring system.

Explainability depth was classified as E1, representing basic or descriptive interpretability outputs; E2, representing explicit intrinsic or post-hoc explanation methods; and E3, representing explanations that were additionally evaluated in terms of technical quality or human-centered relevance, such as faithfulness, stability, comprehensibility, or decision impact [13].

Validation strength was classified as V1 for internal validation within the development dataset or population; V2 for temporal validation or independent testing with limited setting independence; and V3 for independent external, multicenter, geographic, or prospective validation. This classification reflects the increasing strength of evidence for model generalizability emphasized in contemporary prediction-model evaluation guidance [19].

Clinical integration was classified as C1 for research-stage or prototype systems without intended-user evaluation; C2 for clinician-facing, human-in-the-loop, usability, simulated workflow, or early clinical

evaluation; and C3 for prospective integration into real clinical workflows with assessment of clinical decisions, implementation outcomes, patient outcomes, or post-deployment performance [25].

Each study was represented as an E-V-C profile rather than a single composite score. The three dimensions were treated as noncompensatory because strength in one dimension was not assumed to compensate for weakness in another. Thus, advanced explainability cannot substitute for insufficient external validation, and strong validation does not demonstrate clinical integration in the absence of workflow or intended-user evaluation. The resulting profile should therefore be interpreted as a review-specific descriptive instrument rather than a validated clinical-readiness score.

Table 2. Operational Criteria for the Translational Maturity Profile

Dimension	Level	Operational Criterion
Explainability Depth	E1: Foundational	Explanation is limited to basic feature importance, coefficients, attention visualization, or other descriptive outputs without systematic local and global interpretation.
	E2: Intermediate	Explicit post-hoc or intrinsic explanation method is applied, such as SHAP, LIME, EBM, nomogram, counterfactual explanation, or comparable methods, with interpretable local and/or global outputs.
	E3: Advanced	Explanations are additionally evaluated for properties such as faithfulness, robustness, stability, comprehensibility, clinician interpretation, or decision impact.
Validation Strength	V1: Foundational	Evaluation is limited to internal validation, such as train-test split, hold-out testing, bootstrap, or cross-validation within the original dataset or development population.
	V2: Intermediate	Validation extends beyond standard internal resampling, for example temporal validation, independent testing within the same health system, or validation on a distinct dataset with limited population or setting independence.
	V3: Advanced	Independent external validation is conducted across a distinct institution, geographic population, or multicenter cohort, or prospective validation is performed under clinically relevant conditions.
Clinical Integration	C1: Foundational	Model remains at research or prototype stage without evaluation involving intended clinical users or workflow integration.
	C2: Intermediate	System is evaluated through clinician-facing interfaces, human-in-the-loop testing, simulated workflow, usability assessment, or experimental clinical decision-support evaluation.
	C3: Advanced	System is prospectively integrated into real clinical workflow, with evaluation of clinical decision-making, implementation outcomes, patient outcomes, or post-deployment performance.

The ordinal levels indicate increasing depth within each dimension but should not be interpreted as equidistant numerical intervals. Thus, the transition from E1 to E2 is not assumed to represent the same magnitude of advancement as the transition from E2 to E3, and analogous interpretations apply to the validation and clinical-integration dimensions.

3. RESULTS

3.1. Study Selection

A total of 292 records were identified through searches of five databases. After removing duplicates and publications that did not meet the initial criteria, 229 articles entered the title screening stage. At this stage, 212 articles were eliminated because they were not relevant to the domain of mental health, did not involve artificial intelligence methods, or did not implement explainability mechanisms. Three additional articles were excluded during the abstract screening stage, leaving 14 articles to proceed to full-text evaluation. One article was not fully accessible, leaving 13 articles to be assessed for eligibility. Four articles were excluded for not demonstrating clear explainability implementation or lacking sufficient clinical translation relevance, and nine met all inclusion criteria for further analysis. Quantitative details of the selection process are presented in Figure 1. The number of studies that met the inclusion criteria was limited and reflected the current state of the literature, in which the application of XAI that integrates explicit explainability, strong methodological validation, and clinical translation context simultaneously is still rarely reported, indicating that the maturity of XAI translation in mental health is in its early stages of development.

3.2. Study Characteristics

The general characteristics of the nine studies that met the inclusion criteria are summarized in Table 3, covering the year of publication, mental health domain, AI model category, explainability technique, dataset size, validation level, and clinical implementation context. All studies were published after 2020, with an increase in the 2024–2025 period, indicating that the integration of explainability in mental health AI systems is still a relatively new field. In terms of domain, the research focus was dominated by depression and cognitive disorders, while suicide risk and anxiety appeared in more limited proportions. The majority of studies used medium-sized datasets and applied internal validation, with only a small portion involving independent cohorts or external validation strategies. The implementation context was generally a risk prediction or diagnosis system positioned as clinical decision support at the prototype stage.

Table 3. Characteristics of Included Studies

Study	Mental Health Domain	Model Type	XAI Type	Sample Size	Validation Strategy	Deployment Stage
Banerjee et al., 2021 [27]	Severe mental illness	Hybrid ML	Counterfactual (post-hoc)	1,706	Internal	Research stage
Chun et al. (2022) [28]	Dementia conversion	Ensemble ML	SHAP (post-hoc)	705	Internal	Prototype CDSS
Oliveira et al. 3(2024) [29]	Suicide prevention	LLM-based model	LIME (post-hoc)	78	Human-subject evaluation	Human-in-the-loop prototype
de Arriba-Pérez (2025) [30]	Cognitive decline	Hybrid LLM + ML	Feature-based dashboard	44	Internal	Prototype CDSS
Wang et al. (2025) [31]	Perinatal distress	Classical ML (glassbox)	EBM (intrinsic) + SHAP	293	Internal	Prototype clinical support
Zuo & Yang (2025) [32]	Post-stroke depression	Classical ML	SHAP (post-hoc)	1,143	Internal	Web-based prototype
Chen et al. (2025) [33]	Depression risk	Classical ML	SHAP + Nomogram	5,204 + 1,943	Internal + Temporal	Risk prediction tool
Zhang et al., 2025 [34]	Vascular depression	Ensemble ML	SHAP (post-hoc)	602 + 171	External validation	Advanced prototype
Ebrahimi et al. (2026) [35]	Anxiety in ICD patients	Ensemble ML	SHAP (post-hoc)	478	Internal	Remote monitoring prototype

3.3. Distribution of Mental Health Domains

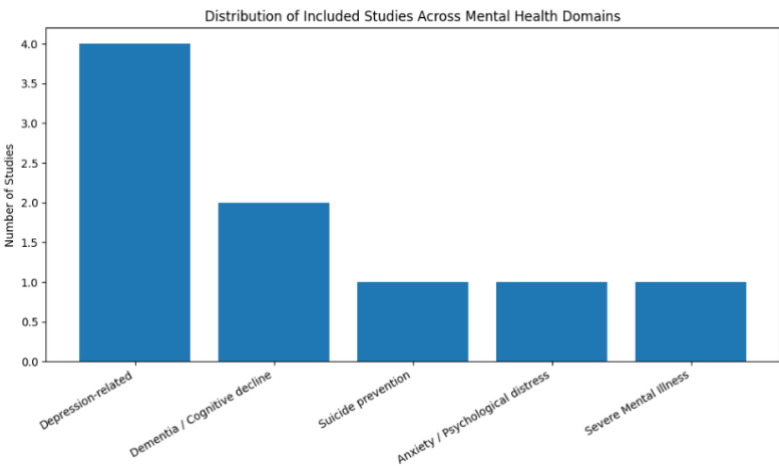


Fig. 2. Distribution of included studies across mental health domains.

The distribution of mental health domains across nine studies shows a primary concentration on the spectrum of depressive disorders, including general depression, post-stroke depression, vascular depression, and perinatal distress, as shown in Figure 2. The next focus is on cognitive disorders, particularly the prediction

of conversion to dementia. Conversely, suicide risk, anxiety or psychological distress in the context of medical comorbidity, and severe mental illness appear in limited proportions in the literature that meets translation criteria. This pattern indicates that the application of XAI in clinical decision support systems is developing primarily in conditions with more structured outcomes and relatively stable data cohorts. Domains with high clinical risk and more complex contextual dynamics, such as suicide risk and severe mental illness, remain underrepresented, reflecting that the development of XAI in mental health is currently more influenced by data availability and maturity than by clinical urgency alone.

3.4. AI Models Used

Analysis of nine studies shows the dominance of classical machine learning approaches over pure deep learning, as summarized in Table 4. Six studies applied tree-based or ensemble algorithms such as Random Forest, XGBoost, and LightGBM, while three studies adopted deep learning architectures or large language models in hybrid configurations. Single models, including Logistic Regression and SVM, were generally used as baseline comparators. Several studies implemented boosting or stacking strategies to improve predictive performance and capture nonlinear interactions between features. In terms of evaluation, eight studies used internal validation through train-test split or cross-validation, while one study reported external validation on an independent cohort. There were no reports of geographical validation or systematic multicenter approaches. Model development was oriented toward internal performance optimization and predictive architecture exploration, with limited testing of generalization capabilities across clinical populations.

Table 4. Model, Explainability, and Translational Characteristics of Included Studies

Study	Model Type	XAI Method	Validation	External	Human Eval	Stage
[27]	Hybrid ML	Counterfactual	Internal CV	No	No	Research
[28]	Classical ML	SHAP	CV + Hold-out	No	No	CDS Prototype
[29]	LLM (NLP)	LIME	Experimental	No	Yes	Human-in-loop
[30]	Hybrid LLM + ML	Dashboard	Internal	No	No	CDSS Prototype
[31]	Classical ML (EBM)	Intrinsic (EBM)	10-fold CV	No	No	Prototype
[32]	Classical ML	SHAP	Split + CV	No	No	Web Prototype
[33]	Classical ML	SHAP + Nomogram	Internal + Temporal	No	No	Risk Tool
[34]	Ensemble ML	SHAP	CV + External	Yes	No	Advanced Prototype
[35]	Ensemble ML	SHAP	5-fold CV	No	No	Remote Monitoring

3.5. Explainability Techniques

The explainability approach in the nine studies analyzed was dominated by post-hoc methods, as summarized in Table 5. Most studies applied Shapley value-based feature attribution techniques to generate global and local explanations for ensemble models or non-linear architectures, while local surrogate approaches such as LIME were used primarily for text-based models or large language models. Several studies chose intrinsically more interpretable models, including Explainable Boosting Machine (EBM) and logistic regression presented in the form of clinical nomograms, while counterfactual or class-contrastive approaches were found in only one study. Reporting of results generally included a combination of global and local explanations, although assessment of clinical relevance was still limited. The involvement of mental health professionals in evaluating explainability was reported in only one study, while other studies tended to present interpretations in the form of technical visualizations without systematic testing of their influence on clinical decision-making, indicating a primary orientation toward computational transparency with minimal integration of clinical meaning.

Table 5. Taxonomy of Explainability Approaches Across Included Studies

XAI Category	Specific Method	Explainability Type	Studies
Shapley-based attribution	SHAP	Post-hoc	[28],[34],[31],[32],[33],[35]
Local surrogate explanation	LIME	Post-hoc	[29],[31]
Intrinsic glassbox modeling	EBM	Intrinsic	[31]
Intrinsic clinical scoring	Nomogram (LR)	Intrinsic	[33]
Counterfactual reasoning	Class-contrastive explanation	Post-hoc	[27]
Custom explanation interface	Feature-based dashboard	Post-hoc	[30]

3.6. Validation and Methodological Quality

The evaluation shows that most studies use internal validation through train–test split or crossvalidation, while independent external validation is only reported in one multicenter study. Temporal validation in the same cohort is still rare, and there is no prospective application in clinical workflows or patient-outcome-based evaluations. This pattern confirms that model testing focuses on performance stability in the original dataset, rather than resilience to population variation, service context, or temporal dynamics, so that the absence of cross-center testing and prospective design limits the assessment of actual clinical utility, especially in mental health, which is characterized by heterogeneity of symptoms and sociodemographic factors.

Methodologically, the majority of studies applied retrospective designs based on secondary datasets with varying sample sizes. Reporting on model calibration, missing data handling, class imbalance, and overfitting mitigation strategies was inconsistent, indicating variation in analytical transparency. Although explainability techniques are reported as part of the model architecture or post-hoc analysis, limitations in validation power and design rigor remain major obstacles in assessing the level of clinical translation maturity, with a summary of PROBAST-based bias risks presented in the next subsection as a systematic evaluation of these methodological aspects.

3.7. Risk of Bias Assessment

Methodological quality assessment was conducted with reference to the PROBAST domain framework, which includes participants, predictors, outcomes, and analysis. This framework serves as a structured evaluation instrument to assess the methodological strength of studies, particularly in relation to validation design, potential overfitting, and generalization capacity directly related to clinical translation readiness. The results of each domain assessment are synthesized into an overall risk estimate at the study level in Table 6, enabling systematic interpretation of the level of methodological maturity within the translational maturity framework used.

In general, the most notable limitations are found in the analysis domain, which is affected by the dominance of internal validation without independent external testing, limited calibration reporting, and the risk of overfitting in single-center retrospective designs with small sample sizes. The risk of bias in the participants domain is moderate due to limitations in population representativeness, while the predictors and outcome domains are relatively more stable despite variations in reporting. One multicenter study with independent external validation showed a lower risk profile, reinforcing the finding that advances in technical explainability have not been consistently followed by adequate validation rigor to support clinical translation maturity.

Table 6. PROBAST Risk-of-Bias Assessment of Included Studies

Study	Participants	Predictors	Outcome	Analysis	Overall Risk
[27]	Unclear	Unclear	Low	High	High
[28]	Unclear	Low	Unclear	High	High
[29]	High	Unclear	Unclear	High	High
[30]	High	Unclear	Unclear	High	High
[31]	Unclear	Low	Unclear	High	High
[32]	Unclear	Low	Unclear	High	High
[33]	Unclear	Low	Unclear	High*	High
[34]	Low	Unclear	Low	Unclear	Unclear
[35]	Unclear	Low	Unclear	High	High

3.8. Translational Maturity Profiles

Application of the predefined E-V-C criteria revealed heterogeneous translational profiles across the nine included studies. As summarized in Table 7, explainability was generally more developed than methodological validation and clinical integration. Most studies reached E2 because they implemented explicit post-hoc or intrinsic explainability methods, particularly SHAP, LIME, counterfactual explanations, or interpretable model structures. One study reached E3 because explainability was evaluated directly with mental health professionals in a human-subject setting.

Validation strength remained predominantly at V1, reflecting reliance on internal cross-validation, hold-out testing, or split-sample evaluation. One study reached V2 through temporal validation, whereas one study reached V3 through independent external validation. Clinical integration showed a similar concentration at the foundational level. Most systems remained research-stage or prototype implementations without direct evaluation involving intended clinical users. One study reached C2 because it incorporated human-in-the-loop

evaluation with mental health professionals. No included study met the predefined criteria for C3, which required prospective integration into routine clinical workflow with evaluation of clinical, implementation, or patient-level outcomes.

These profiles indicate an asymmetric translational pattern within the selected evidence base. The presence of technically explicit explainability mechanisms was not consistently accompanied by stronger external validation or clinical workflow integration. Importantly, the E-V-C profiles are descriptive and dimension-specific; they do not represent a composite maturity score or a validated measure of clinical readiness.

Table 7. Review-Specific Translational Maturity Profiles of Included Studies

Study	Explainability	Validation	Clinical Integration	E-V-C Profile	Main basis for classification
Banerjee et al. [27]	E2	V1	C1	E2-V1-C1	Counterfactual explanation; internal validation; research-stage application
Chun et al. [28]	E2	V1	C1	E2-V1-C1	SHAP-based explanation; internal CV/hold-out; CDSS prototype
de Arriba-Pérez et al. [30]	E2	V1	C1	E2-V1-C1	Interpretable dashboard; internal evaluation; prototype CDSS
Wang et al. [31]	E2	V1	C1	E2-V1-C1	Intrinsic EBM with SHAP; 10-fold CV; clinical-support prototype
Zuo and Yang [32]	E2	V1	C1	E2-V1-C1	SHAP explanation; split-sample/CV; web-based prototype
Chen et al. [33]	E2	V2	C1	E2-V2-C1	SHAP and nomogram; temporal validation; risk-prediction tool
Zhang et al. [34]	E2	V3	C1	E2-V3-C1	SHAP explanation; independent external validation; advanced prototype without reported user evaluation
Ebrahimi et al. [35]	E2	V1	C1	E2-V1-C1	SHAP explanation; 5-fold CV; remote-monitoring prototype

4. DISCUSSION

4.1. Synthesis of Key Findings

Within the broader framework of biomedical decision systems, empirical studies that meet translation criteria during the review period have almost exclusively focused on mental health. This pattern indicates that the fields of psychiatry and cognitive disorders are currently the primary arena for translational XAI exploration [12]. The high-risk, contextual, and multidimensional nature of decision-making makes explainability a clinical prerequisite, not merely a technical feature [36], thus mental health represents a high-risk domain relevant for testing translation challenges in biomedical decision systems in general.

Methodological synthesis shows the dominance of SHAP as an interpretation mechanism in ensemble and boosting models for diagnosis and risk prediction [28], [34], including in the development of cohort-based depression models and psychological monitoring systems [33], [35]. LIME is more commonly used in language-based models or large language models [29], while the counterfactual approach through class-contrastive reasoning has only been reported in severe mental illness mortality prediction [27]. This variation indicates that the selection of explainability techniques is determined more by model architecture than by substantive clinical needs.

Clinically, XAI applications are concentrated on conditions with measurable outcomes and relatively stable cohort datasets, particularly various forms of post-stroke depression [32], vascular depression in the elderly population [34], and perinatal distress [31]. Predictions of cognitive impairment conversion to dementia have also received significant attention [28], [30]. Suicide prevention and anxiety in the context of medical comorbidity are relatively underrepresented [29], [35], reflecting a reliance on more standardized data structures. Validation practices are dominated by internal evaluation through crossvalidation or train-test split [27], [28]. Independent external validation is only reported in the vascular depression diagnostic model, which is also compared with medical personnel [34], while temporal validation is applied in the same national cohort context [33]. The absence of prospective implementation in clinical workflows places the system at the prototype stage, consistent with PROCAST findings indicating a high risk of bias in the analysis domain due

to potential overfitting and the absence of cross-population testing, thus limiting translation due to methodological designs that are not yet adequate for widespread clinical adoption.

4.2. Explainability: Epistemology and Human-Centered Implications

The included studies predominantly relied on post-hoc explainability methods, particularly SHAP, with more limited use of LIME, intrinsic models, and counterfactual explanations. This distribution is important because these approaches provide different forms of information and should not be interpreted as methodologically equivalent. SHAP assigns feature contributions to model predictions relative to a specified reference or background distribution and can provide useful global and local summaries of model behavior. However, these attributions primarily describe how the trained model uses available predictors and does not, by themselves, establish that the identified features represent causal determinants of the clinical outcome. In correlated and heterogeneous mental health data, attribution patterns may therefore reflect statistical dependencies embedded in the dataset and model rather than underlying disease mechanisms [13].

LIME provides a different form of post-hoc explanation by approximating the behavior of a complex model with an interpretable surrogate in the local neighborhood of a particular prediction. Its local nature can facilitate case-level interpretation, particularly for text-based or high-dimensional models. Nevertheless, the resulting explanation depends on how the local neighborhood is generated and weighted, meaning that the surrogate explanation represents an approximation of local model behavior rather than a direct description of the internal decision process of the original model [37]. These characteristics are particularly relevant in clinical applications, where apparently intuitive explanations may be interpreted more strongly than their methodological assumptions justify.

Intrinsically interpretable models provide a different epistemic position because interpretability is embedded in the predictive structure rather than reconstructed after prediction. Models such as logistic regression, nomograms, and explainable boosting machines can expose predictor-response relationships more directly. In high-stakes applications, this structural transparency may offer advantages over post-hoc approximation when comparable predictive performance can be achieved [38]. Intrinsic interpretability should not itself be equated with causal validity. A transparent model can still learn non-causal associations, reflect confounding, or depend on biased predictor selection.

Counterfactual explanations offer an additional perspective by describing how changes in selected features could alter a model prediction. This form of explanation can be clinically intuitive because it moves beyond feature importance toward contrastive reasoning. However, a model-generated counterfactual should not automatically be interpreted as an actionable clinical intervention. A feature change that alters a prediction mathematically may be biologically implausible, clinically infeasible, or causally ineffective if dependencies among variables are not adequately represented.

These distinctions highlight an important boundary between model explanation and causal or mechanistic explanation. Model explanation addresses why a particular computational model produced a prediction given its learned statistical structure. In contrast, causal or mechanistic explanation concerns why the clinical phenomenon occurs and whether changing a factor would alter the outcome through a plausible causal pathway. The two forms of explanation may overlap, but they are not equivalent. Consequently, feature importance, SHAP values, local surrogate explanations, or counterfactual outputs should not be interpreted as evidence of clinical causation unless the underlying study design and analytical assumptions explicitly support causal inference.

This distinction is particularly important in mental health, where predictors often include interacting behavioral, sociodemographic, physiological, and contextual variables. In this setting, technically transparent model outputs may improve inspectability without necessarily providing clinically valid explanations. The findings of this review therefore support evaluating XAI not only in terms of whether an explanation is produced, but also in terms of its fidelity to the model, stability, clinical comprehensibility, and influence on decision-making.

4.3. Translational and Governance Gaps

Limitations in validation and the lack of prospective implementation indicate a clear translational gap. Most systems are still in the prototype or experimental development phase, despite being positioned as clinical

decision support [32], [35]. Independent external validation was found in only one multicenter study, while evaluation of patient outcomes or long-term impact was not reported [34].

Reliance on a single dataset and the dominance of internal validation increase the risk of optimism bias, especially in complex architectures or small sample sizes [30], [29]. Under these conditions, feature attribution visualization through SHAP and similar methods has the potential to create an explainability illusion, namely the assumption that algorithmic transparency reflects clinical readiness without adequate evidence of generalization. This risk is particularly relevant in mental health, which is characterized by symptom heterogeneity and sociodemographic variation.

The limited use of independent external validation and the absence of prospective workflow implementation indicate that the translational evidence within the included mental health studies remains constrained. Most systems were evaluated at the research or prototype stage, and only a small proportion incorporated user-facing evaluation or independent validation. These findings should be interpreted specifically within the evidence base included in this review and should not be generalized to the maturity of XAI across other biomedical domains.

Mental health nevertheless represents a particularly demanding context for XAI translation because clinical decisions are influenced by heterogeneous symptoms, contextual factors, sociodemographic variation, and uncertainty in diagnostic boundaries. In such settings, technical explainability alone is insufficient to demonstrate clinical readiness. Robust validation, meaningful interaction with intended users, and evaluation within actual clinical workflows remain necessary before explainable models can be considered suitable for routine decision support.

4.4. Translational Maturity Profiles Across Explainability, Validation, and Clinical Integration

The readiness of XAI translation in mental health decision systems is best understood as an interaction between three interrelated dimensions: depth of explainability, strength of methodological validation, and level of clinical integration. Unlike translation models that emphasize a gradual progression from innovation to implementation, this framework places clinical readiness as a multidimensional construct. Structural transparency, robustness of generalization, and embeddedness in clinical workflows do not develop automatically or simultaneously. This three-dimensional approach fills a conceptual void in translation models by positioning explainability as an evaluative variable on par with validation and clinical integration, rather than merely an additional attribute.

 Explainability Depth	 Validation Strength	 Clinical Integration
Advanced - Explanation empirically evaluated with clinicians - Demonstrated influence on decision interpretation or trust	Advanced - Independent external cohort - Multicenter or prospective validation	Advanced - Integrated within real-world clinical workflow - Operational deployment with outcome monitoring
Intermediate - Local and global explanations provided (SHAP, LIME) - No formal evaluation of clinical impact	Intermediate - Temporal validation - Hold-out testing within same dataset	Intermediate - Human-in-the-loop testing - Simulated or experimental workflow evaluation
Foundational - Feature importance only - No structured interpretability strategy	Foundational Internal cross-validation only	Foundational - Research-stage prototype - No real-world evaluation

Fig. 3. Operational Translational Maturity Framework

Depth of explainability refers to the structural nature of interpretability mechanisms, ranging from intrinsically transparent models to post-hoc approaches based on feature attribution and counterfactual reasoning. Post-hoc approaches such as SHAP dominate ensemble and boosting models [28], [34], [35], while structurally transparent intrinsic models, such as EBM [31] or logistic regressionbased nomograms, are

relatively rarely used [33]. This pattern indicates that most studies are at a medium level of technical explainability. Direct clinical evaluation or user-based testing of is rarely conducted, with limited exceptions in studies testing mental health professionals' trust [29].

In terms of validation, the majority of studies still rely on internal evaluation through cross-validation or split-sample [27], [28], [35], while temporal validation in the same cohort is only reported to a limited extent [33], and independent multicenter external validation is only found in one study [34]. Clinical integration is generally limited to clinical decision support prototypes or exploratory dashboards [32], [35], without prospective implementation and without patient outcome evaluation. This dominant configuration places most studies at an early stage of translational maturity, as shown in Figure 3, where computational transparency has been achieved to a certain extent, but has not yet been matched by robustness of generalization or systematic clinical embedding. Translational maturity of XAI is the result of a dynamic interaction between explanation structure, methodological strength, and operational implementation readiness.

5. LIMITATIONS

This structured review includes only nine studies that met the predefined eligibility criteria. The limited evidence base restricts the breadth of interpretation and requires the observed patterns to be understood as characteristics of the selected literature rather than as a complete representation of XAI research in mental health.

The review included only peer-reviewed English-language journal articles. This criterion supported consistent extraction of methodological, validation, and clinical information, but it may have excluded relevant methodological developments reported in conference proceedings or non-English publications. This limitation is particularly relevant to artificial intelligence research, where technical innovations are frequently disseminated through conference venues.

A single reviewer conducted study screening, eligibility assessment, data extraction, PROBABST appraisal, and E-V-C classification. Predefined criteria, structured extraction procedures, and repeated verification against the full-text articles were used to maintain consistency and traceability. However, these procedures do not replace independent duplicate assessment and cannot eliminate the possibility of selection or interpretation bias. Inter-rater reliability was therefore not available for the review process.

The E-V-C translational maturity profile represents a descriptive instrument developed specifically for this review. The operational criteria provide a structured basis for comparing explainability depth, validation strength, and clinical integration across the included studies. However, the framework has not undergone independent reliability testing, construct validation, or prospective clinical validation. The resulting profiles should therefore not be interpreted as validated clinical-readiness scores or deployment thresholds.

The methodological appraisal also depended on the completeness of reporting in the primary studies. Incomplete descriptions of calibration, missing-data handling, analytical procedures, explanation evaluation, or workflow integration may have limited the precision of several judgments. Future studies should evaluate the reproducibility of the E-V-C profile using larger evidence bases and independent reviewers.

6. FUTURE DIRECTIONS

Future research on XAI in mental health should extend evaluation beyond predictive performance and the mere availability of explanations. Model development should incorporate assessment of explanation fidelity, stability, robustness, and comprehensibility alongside conventional measures of discrimination and calibration. External, temporal, or multicenter validation should also precede claims regarding clinical generalizability.

Clinical evaluation should involve intended users at an earlier stage of system development. Human-in-the-loop studies can examine whether explanations improve clinician understanding, influence clinical decisions appropriately, reduce or increase cognitive burden, and affect reliance on AI-generated recommendations. Prospective evaluation within real clinical workflows remains necessary to determine whether technically interpretable models provide meaningful support under routine service conditions.

Healthcare organizations should evaluate model performance, explainability, population generalizability, workflow compatibility, and post-deployment behavior as related but distinct requirements. Local validation can identify performance degradation when systems are transferred to populations or settings that differ from the original development environment. Continuous monitoring can also detect model drift, changes in explanation patterns, and unintended reliance after implementation.

Governance and regulatory evaluation should distinguish technical explainability from evidence of clinical readiness. The presence of SHAP values, feature-importance visualizations, or other explanation mechanisms does not independently demonstrate safety, generalizability, or clinical utility. Evaluation

standards should therefore consider the quality of explanations together with external validation, human-centered assessment, workflow integration, and post-deployment monitoring.

Future reviews may also examine the reliability and validity of the E-V-C profile across broader evidence bases and different clinical contexts. Independent coding and comparison across reviewers would provide evidence regarding the reproducibility of the proposed classification structure.

7. CONCLUSION

This structured review indicates that the nine included studies demonstrate greater development in computational explainability than in methodological validation and clinical integration. Post-hoc feature-attribution methods, particularly SHAP, dominate the selected literature, whereas intrinsic and counterfactual approaches appear less frequently. Most studies rely on internal validation, while independent external validation, direct evaluation with intended users, and prospective clinical workflow integration remain limited.

The findings also demonstrate that model explainability cannot be treated as equivalent to clinical readiness. An explanation may clarify how a computational model generates a prediction, but it does not independently establish model generalizability, causal validity, clinical usefulness, or safe integration into healthcare workflows.

The E-V-C profile provides a structured approach for describing explainability depth, validation strength, and clinical integration without combining these dimensions into a single composite score. The profile remains a review-specific descriptive instrument and has not been validated as a clinical-readiness measure.

The interpretation of these findings remains constrained by the small evidence base, journal-only selection, single-reviewer methodology, and dependence on the reporting quality of the primary studies. The conclusions therefore apply primarily to the studies included in this review and should not be generalized directly to the overall maturity of XAI in mental health or other biomedical domains. Further progress toward clinical translation requires stronger external validation, human-centered explanation evaluation, prospective workflow testing, and post-deployment monitoring.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to improve language clarity and readability. Scopus AI was used to support comparison and cross-checking of results obtained from the manually conducted literature search. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

- [1] N. Koutsouleris, T. U. Hauser, V. Skvortsova, and M. De Choudhury, "From promise to practice: towards the realisation of AI-informed mental health care," *Lancet Digit. Health*, vol. 4, no. 11, pp. e829–e840, Nov. 2022, doi: [10.1016/S2589-7500\(22\)00153-4](https://doi.org/10.1016/S2589-7500(22)00153-4).
- [2] G. Antoniou, E. Papadakis, and G. Baryannis, "Mental Health Diagnosis: A Case for Explainable Artificial Intelligence," *International Journal on Artificial Intelligence Tools*, vol. 31, no. 03, May 2022, doi: [10.1142/S0218213022410032](https://doi.org/10.1142/S0218213022410032).
- [3] Y. Hong and Z. Xia, "AI-Driven Innovations in Psychological Assessment: Multimodal Data, Intelligent Analytics, and Ethical Challenges," in *Proceedings of the 2025 International Conference on Artificial Intelligence and Smart Manufacturing*, New York, NY, USA: ACM, May 2025, pp. 854–859. doi: [10.1145/3756423.3756564](https://doi.org/10.1145/3756423.3756564).
- [4] D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *NPJ Digit. Med.*, vol. 6, no. 1, p. 6, Jan. 2023, doi: [10.1038/s41746-023-00751-9](https://doi.org/10.1038/s41746-023-00751-9).
- [5] I. Levkovich, S. Shinan-Altman, and Z. Elyoseph, "Can large language models be sensitive to culture suicide risk assessment?," *J. Cult. Cogn. Sci.*, vol. 8, no. 3, pp. 275–287, Dec. 2024, doi: [10.1007/s41809-024-00151-9](https://doi.org/10.1007/s41809-024-00151-9).
- [6] I. Ahmed, A. Brahmacharimayum, R. H. Ali, T. A. Khan, and M. O. Ahmad, "Explainable AI for Depression Detection and Severity Classification From Activity Data: Development and Evaluation Study of an Interpretable Framework," *JMIR Ment. Health*, vol. 12, pp. e72038–e72038, Sep. 2025, doi: [10.2196/72038](https://doi.org/10.2196/72038).
- [7] X. Li and M. Mahmoud, "Unlocking the Black Box: Concept-Based Modeling for Interpretable Affective Computing Applications," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, May 2024, pp. 1–10. doi: [10.1109/FG59268.2024.10581918](https://doi.org/10.1109/FG59268.2024.10581918).
- [8] A. Malhotra and R. Jindal, "XAI Transformer based Approach for Interpreting Depressed and Suicidal User Behavior on Online Social Networks," *Cogn. Syst. Res.*, vol. 84, p. 101186, Mar. 2024, doi: [10.1016/j.cogsys.2023.101186](https://doi.org/10.1016/j.cogsys.2023.101186).
- [9] C.-C. Tsai *et al.*, "Effect of Artificial Intelligence Helpfulness and Uncertainty on Cognitive Interactions with Pharmacists: Randomized Controlled Trial," *J. Med. Internet Res.*, vol. 27, p. e59946, Jan. 2025, doi: [10.2196/59946](https://doi.org/10.2196/59946).

- [10] A. Gerdes, "The role of explainability in AI-supported medical decision-making," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 29, Apr. 2024, doi: [10.1007/s44163-024-00119-2](https://doi.org/10.1007/s44163-024-00119-2).
- [11] H. Liu, "Interpretable Artificial Intelligence (XAI) in Biomedicine: Status, Challenges and Interdisciplinary Solutions," *Theoretical and Natural Science*, vol. 117, no. 1, pp. 95–101, Jul. 2025, doi: [10.54254/2753-8818/2025.LD25347](https://doi.org/10.54254/2753-8818/2025.LD25347).
- [12] Suneel Pappala, M. Malyadri, K. Venkata Naganjaneyulu, A. Prakashini, and Pasuladi Santosh, "Explainable AI for healthcare professionals: Advancing risk assessment, diagnostic precision, and ethical clinical interventions," *World Journal of Biology Pharmacy and Health Sciences*, vol. 22, no. 1, pp. 446–453, Apr. 2025, doi: [10.30574/wjbphs.2025.22.1.0425](https://doi.org/10.30574/wjbphs.2025.22.1.0425).
- [13] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Inform.*, vol. 113, p. 103655, Jan. 2021, doi: [10.1016/j.jbi.2020.103655](https://doi.org/10.1016/j.jbi.2020.103655).
- [14] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, p. 107161, Nov. 2022, doi: [10.1016/j.cmpb.2022.107161](https://doi.org/10.1016/j.cmpb.2022.107161).
- [15] J. Jung, H. Lee, H. Jung, and H. Kim, "Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review," *Heliyon*, vol. 9, no. 5, p. e16110, May 2023, doi: [10.1016/j.heliyon.2023.e16110](https://doi.org/10.1016/j.heliyon.2023.e16110).
- [16] T. Greenhalgh *et al.*, "Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies," *J. Med. Internet Res.*, vol. 19, no. 11, p. e367, Nov. 2017, doi: [10.2196/jmir.8775](https://doi.org/10.2196/jmir.8775).
- [17] X. Liu *et al.*, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," *Nat. Med.*, vol. 26, no. 9, pp. 1364–1374, Sep. 2020, doi: [10.1038/s41591-020-1034-x](https://doi.org/10.1038/s41591-020-1034-x).
- [18] 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, May 2022, doi: [10.1038/s41591-022-01772-9](https://doi.org/10.1038/s41591-022-01772-9).
- [19] 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, p. e078378, Apr. 2024, doi: [10.1136/bmj-2023-078378](https://doi.org/10.1136/bmj-2023-078378).
- [20] K. Lekadir *et al.*, "FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, vol. 388, p. e081554, Feb. 2025, doi: [10.1136/bmj-2024-081554](https://doi.org/10.1136/bmj-2024-081554).
- [21] A. H. van der Vegt, I. A. Scott, K. Dermawan, R. J. Schnetler, V. R. Kalke, and P. J. Lane, "Implementation frameworks for end-to-end clinical AI: derivation of the SALIENT framework," *Journal of the American Medical Informatics Association*, vol. 30, no. 9, pp. 1503–1515, Aug. 2023, doi: [10.1093/jamia/ocad088](https://doi.org/10.1093/jamia/ocad088).
- [22] N. R. Haddaway, M. J. Page, C. C. Pritchard, and L. A. McGuinness, "PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis," *Campbell Systematic Reviews*, vol. 18, no. 2, Jun. 2022, doi: [10.1002/cl2.1230](https://doi.org/10.1002/cl2.1230).
- [23] S. Cruz Rivera *et al.*, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," *Nat. Med.*, vol. 26, no. 9, pp. 1351–1363, Sep. 2020, doi: [10.1038/s41591-020-1037-7](https://doi.org/10.1038/s41591-020-1037-7).
- [24] X. Liu *et al.*, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," *Nat. Med.*, vol. 26, no. 9, pp. 1364–1374, Sep. 2020, doi: [10.1038/s41591-020-1034-x](https://doi.org/10.1038/s41591-020-1034-x).
- [25] 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, May 2022, doi: [10.1038/s41591-022-01772-9](https://doi.org/10.1038/s41591-022-01772-9).
- [26] R. F. Wolff *et al.*, "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, Jan. 2019, doi: [10.7326/M18-1376](https://doi.org/10.7326/M18-1376).
- [27] S. Banerjee, P. Lio, P. B. Jones, and R. N. Cardinal, "A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness," *NPJ Schizophr.*, vol. 7, no. 1, Dec. 2021, doi: [10.1038/s41537-021-00191-y](https://doi.org/10.1038/s41537-021-00191-y).
- [28] M. Y. Chun *et al.*, "Prediction of conversion to dementia using interpretable machine learning in patients with amnesic mild cognitive impairment," *Front. Aging Neurosci.*, vol. 14, Aug. 2022, doi: [10.3389/fnagi.2022.898940](https://doi.org/10.3389/fnagi.2022.898940).
- [29] A. Caetano *et al.*, "Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. Effect of Explainable Artificial Intelligence on Trust of Mental Health Professionals in an AI-based System for Suicide Prevention", doi: [10.1109/ACCESS.2024.0429000](https://doi.org/10.1109/ACCESS.2024.0429000).
- [30] F. de Arriba-Pérez and S. García-Méndez, "Leveraging large language models through natural language processing to provide interpretable machine learning predictions of mental deterioration in real time," *Arab. J. Sci. Eng.*, vol. 50, no. 15, pp. 11577–11591, Aug. 2025, doi: [10.1007/s13369-024-09508-2](https://doi.org/10.1007/s13369-024-09508-2).
- [31] G. Wang *et al.*, "Investigating Protective and Risk Factors and Predictive Insights for Aboriginal Perinatal Mental Health: Explainable Artificial Intelligence Approach," *J. Med. Internet Res.*, vol. 27, no. 1, 2025, doi: [10.2196/68030](https://doi.org/10.2196/68030).

-
- [32] W. Zuo and X. Yang, "Network-based predictive models for artificial intelligence: an interpretable application of machine learning techniques in the assessment of depression in stroke patients," *BMC Geriatr.*, vol. 25, no. 1, Dec. 2025, doi: [10.1186/s12877-025-05837-5](https://doi.org/10.1186/s12877-025-05837-5).
- [33] J. Chen, Y. Lin, R. Hu, and C. Hu, "Prediction model for depression risk in middle-aged and elderly patients with metabolic syndrome: a nomogram and interpretable machine learning approach based on CHARLS," *BMC Psychiatry*, vol. 25, no. 1, Dec. 2025, doi: [10.1186/s12888-025-07434-7](https://doi.org/10.1186/s12888-025-07434-7).
- [34] R. Zhang *et al.*, "Identification and validation of an explainable machine learning model for vascular depression diagnosis in the older adults: a multicenter cohort study," *BMC Med.*, vol. 23, no. 1, Dec. 2025, doi: [10.1186/s12916-025-04283-9](https://doi.org/10.1186/s12916-025-04283-9).
- [35] A. Ebrahimi *et al.*, "Explainable AI models for identifying anxiety and distress in cardiac patients with ICDs," *BMC Med. Inform. Decis. Mak.*, Dec. 2025, doi: [10.1186/s12911-025-03315-x](https://doi.org/10.1186/s12911-025-03315-x).
- [36] M. Alkan, I. Zakariyya, S. Leighton, K. B. Sivangi, C. Anagnostopoulos, and F. Deligianni, "Artificial Intelligence-Driven Clinical Decision Support Systems," Feb. 2025.
- [37] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?,'" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 1135–1144. doi: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778).
- [38] 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, May 2019, doi: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x).
- [39] S. Reddy, "Explainability and artificial intelligence in medicine," *Lancet Digit. Health*, vol. 4, no. 4, pp. e214–e215, Apr. 2022, doi: [10.1016/S2589-7500\(22\)00029-2](https://doi.org/10.1016/S2589-7500(22)00029-2).
- [40] J. M. Bauer and M. Michalowski, "Human-centered explainability evaluation in clinical decision-making: a critical review of the literature," *Journal of the American Medical Informatics Association*, vol. 32, no. 9, pp. 1477–1484, Sep. 2025, doi: [10.1093/jamia/ocaf110](https://doi.org/10.1093/jamia/ocaf110).
- [41] W. Alharbi and A. A. Alfayez, "Explainable artificial intelligence in pancreatic cancer prediction: from transparency to clinical decision-making," *Front. Oncol.*, vol. 15, Dec. 2025, doi: [10.3389/fonc.2025.1720039](https://doi.org/10.3389/fonc.2025.1720039).
- [42] H. Eshkiki, F. Tanhaei, F. Caraffini, and B. Mora, "A Survey of the Application of Explainable Artificial Intelligence in Biomedical Informatics," *Applied Sciences*, vol. 15, no. 24, p. 12934, Dec. 2025, doi: [10.3390/app152412934](https://doi.org/10.3390/app152412934).
- [43] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, Jan. 2022, doi: [10.1038/s41591-021-01614-0](https://doi.org/10.1038/s41591-021-01614-0).
- [44] A. Nicolson, E. Bradburn, Y. Gal, A. T. Papageorgiou, and J. A. Noble, "The human factor in explainable artificial intelligence: clinician variability in trust, reliance, and performance," *NPJ Digit. Med.*, vol. 8, no. 1, p. 658, Nov. 2025, doi: [10.1038/s41746-025-02023-0](https://doi.org/10.1038/s41746-025-02023-0).
-