Original Article *Journal of Multidisciplinary Research Advancements (JOMRA), Vol. 4, No. 1 (2026), 52 – 61*
e-ISSN 2990-7721 Available online at https://www.nepjol.info/index.php/jomra

Bridging the Trust Gap: A Framework for Explainable AI in Mental Health Diagnostics

Collins Onyemaobi[1]

[1]Department of Artificial Intelligence, IU University of Applied Sciences, Germany

**Corresponding author: ct.onyemaobi@gmail.com*

Abstract: The integration of AI into mental health diagnostics offers potential gains in accuracy, scalability, and personalisation. However, clinical adoption remains limited by a trust deficit rooted in model opacity, algorithmic bias, and misalignment with ethical-legal standards. This paper proposes a conceptual framework for Explainable AI (XAI) specifically designed for mental health applications. Through a critical synthesis of literature, we identify three interconnected trust barriers: the black-box problem, bias propagation from non-representative datasets, and the lack of actionable explanations for clinicians. Our framework addresses these barriers through four integrated pillars: (1) a hybrid XAI strategy combining post-hoc (SHAP, Grad-CAM) and intrinsic (attention) techniques; (2) mandatory evaluation of explanations using faithfulness and stability metrics; (3) embedded, proactive bias mitigation across the AI pipeline; and (4) an ethical-regulatory scaffold compliant with GDPR, HIPAA, and the EU AI Act, while prioritising clinical workflow integration. The framework shifts the paradigm from accuracy-first to trust-first design. It provides a structured blueprint for developing transparent, fair, and clinically actionable AI systems, thereby bridging the explainability gap and laying the conceptual foundation for trustworthy AI-augmented mental health care.

Keywords: *Explainable Artificial Intelligence (XAI), Mental Health Diagnostics, Algorithmic Bias, Trust in AI, Clinical Decision Support, Ethical AI, Healthcare AI, Transparency, Fairness, Regulatory Compliance*

Conflicts of interest: None
Supporting agencies: None

Received 07.02.2026 Revised 19.04.2026 Accepted 24.05.2026

Cite This Article: Onyemaobi, C. (2026). Bridging the Trust Gap: A Framework for Explainable AI in Mental Health Diagnostics. *Journal of Multidisciplinary Research Advancements*, *4*(1), 52-61.

1. Introduction

Artificial Intelligence (AI) is rapidly transforming healthcare, offering unprecedented capabilities in diagnostics, predictive analytics, and personalised treatment strategies (Fahim et al., 2025). Within mental healthcare specifically, AI applications have evolved to include predictive diagnostics, tailored treatment recommendations, and continuous monitoring through natural language processing (NLP) and affective computing (Olawade et al., 2024). These advancements promise to alleviate burdens on healthcare systems and improve access to care. However, despite this significant promise, the widespread adoption of AI in clinical mental health practice remains limited, primarily due to profound challenges related to transparency, trust, and ethical accountability (Lee et al., 2021).

A central barrier is the "black-box" nature of many advanced AI models, particularly deep learning systems (Qamar & Bawany, 2023). These models often operate as opaque predictors, generating diagnostic outputs without providing human-interpretable justifications for their decisions (Morley et al., 2020). This lack of interpretability forces clinicians to rely on model suggestions without understanding the underlying reasoning, creating a significant trust deficit. In high-stakes domains like mental health, where diagnostic decisions have profound personal and social consequences, this opacity is not merely a technical inconvenience but a critical ethical and practical flaw (Kerasidou, 2021). The abstracted depiction of this problematic interaction, in which AI models deliver outputs without explanatory insight to end users, underscores the core of the adoption challenge (Borys et al., 2023).

This trust gap is further widened by the pervasive risk of bias in AI systems. Models trained on non-representative or historically skewed datasets risk perpetuating and amplifying existing healthcare disparities (Norori et al., 2021). In mental health, where subjective assessment and socio-cultural factors are deeply intertwined, such biases can lead to inaccurate diagnoses and inequitable treatment recommendations for underrepresented populations (Binns, 2018). Consequently, an

AI system that is both inscrutable and potentially biased poses a serious threat to fairness and equity in mental healthcare delivery (Timmons et al., 2023).

Explainable AI (XAI) has emerged as a pivotal field that seeks to address these challenges by making AI decision-making processes transparent, interpretable, and accountable. Techniques such as SHAP (SHapley Additive Explanations) and Grad-CAM (Gradient-weighted Class Activation Mapping) offer post-hoc explanations, while intrinsic methods like attention mechanisms build interpretability directly into models (Ribeiro et al., 2016; Borys et al., 2023). Yet, the application of XAI in mental health remains underexplored and often lacks robust frameworks that evaluate not only the presence of an explanation but also its quality. Key quality metrics—faithfulness (whether explanations accurately reflect the model's true reasoning) and stability (whether similar inputs yield consistent explanations)—are essential for ensuring that AI-generated insights align reliably with clinical reasoning (Torous & Blease, 2024).

Therefore, this study argues that closing the trust gap in mental health AI requires more than the ad-hoc application of explainability techniques. It necessitates a dedicated conceptual framework that systematically integrates both post-hoc and intrinsic XAI methodologies, rigorously evaluates them against faithfulness and stability metrics, and explicitly incorporates bias mitigation and ethical governance from the outset. Through a critical review of current limitations and a synthesis of best practices, we propose such a framework. This framework is designed to advance the development of transparent, trustworthy, and fair AI systems, ultimately fostering their responsible integration into mental health diagnostics and supporting the evolution of ethically accountable clinical decision-support tools.

2. Materials and methods

This study employs a conceptual and integrative review methodology to develop a comprehensive framework for Explainable AI (XAI) in mental health diagnostics. The objective is not to present new empirical data, but to synthesise existing literature, identify critical gaps, and propose a structured, principled approach to designing transparent, fair, and clinically actionable AI systems. The methodology is structured around three core pillars: (1) a systematic identification of the trust deficit and its root causes in mental health AI; (2) the synthesis of explainability techniques and evaluation metrics tailored to psychiatric contexts; and (3) the formulation of an integrated framework that binds technical explainability with ethical and regulatory imperatives.

2.1. Analytical Approach and Literature Synthesis

The foundation of this framework is a critical narrative review of contemporary literature at the intersection of AI, mental health, explainability, and ethics. Sources were identified through searches of major academic databases (e.g., IEEE Xplore, PubMed, ACM Digital Library, Google Scholar) using keywords such as "explainable AI," "mental health diagnostics," "algorithmic bias," "trust in clinical AI," and "regulatory compliance." Priority was given to recent peer-reviewed articles (2018–2024), systematic reviews, and influential policy documents. The analysis focused on:

•Technical Limitations: The "black-box" problem of deep learning models and the insufficiency of performance metrics (accuracy, AUROC) alone for clinical adoption (Morley et al., 2020; Lee et al., 2021).

•Clinical and Ethical Challenges: Issues of bias propagation from non-representative datasets, fairness in outcomes, and the erosion of clinician and patient trust (Norori et al., 2021; Kerasidou, 2021).

•Existing XAI Solutions: A review of both intrinsic (e.g., attention mechanisms, interpretable models) and post-hoc (e.g., SHAP, LIME, Grad-CAM) techniques, assessing their strengths, limitations, and documented applications in healthcare (Ribeiro et al., 2016; Borys et al., 2023).

This synthesis revealed a significant gap: while XAI tools exist, their application in mental health is often piecemeal, lacking a unified approach that simultaneously addresses faithfulness, stability, bias mitigation, and clinical workflow integration.

2.2. Framework Development: Core Components

To address these gaps, we propose a conceptual XAI framework built on four interdependent components, visualised in Figure 1.

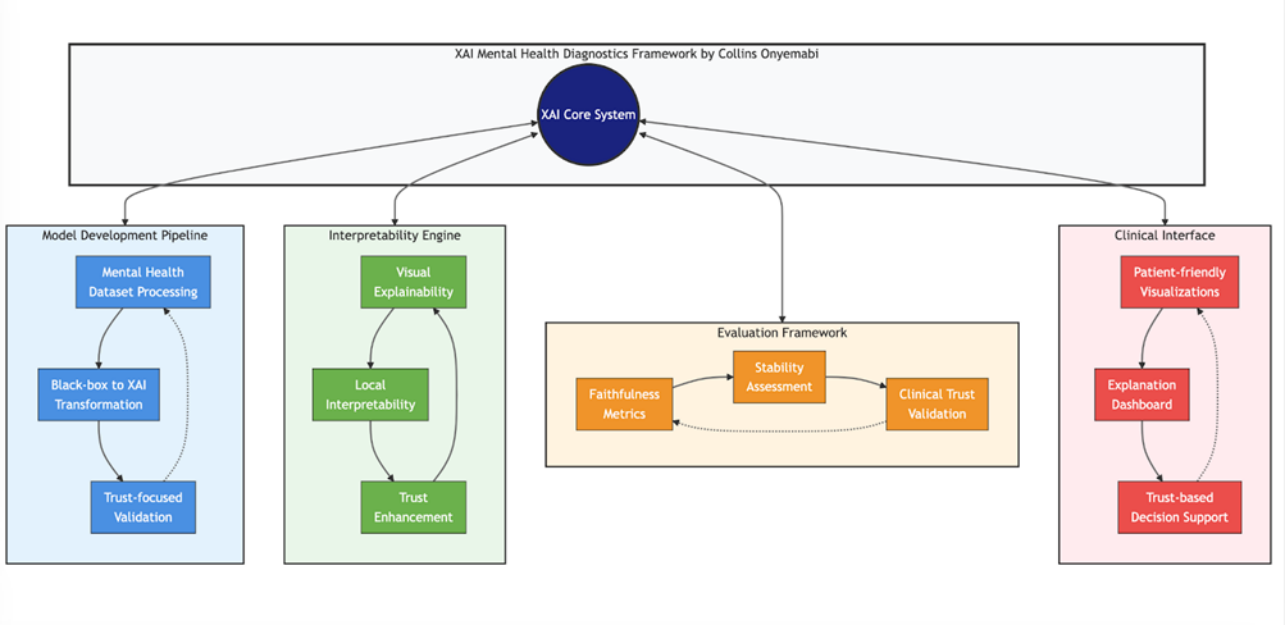

Figure 1: The XAI Mental Health Diagnostics Framework

Hybrid Explainability Integration

The framework advocates for a dual-strategy approach that combines:

• Intrinsic Explainability: Incorporating interpretability directly into model architectures. This includes using attention mechanisms in NLP models to highlight clinically salient words or phrases in patient narratives, and designing neural networks with built-in feature attribution layers to ensure stability and consistency from the outset.

• Post-hoc Explainability: Applying robust explanation techniques after model training. The framework prioritises SHAP for structured clinical data due to its theoretical grounding in game theory and consistent feature attribution, and Grad-CAM for imaging data to provide visual heatmaps that localise diagnostic reasoning to anatomically or clinically relevant regions (Borys et al., 2023). Techniques such as LIME were deprioritised in this conceptualisation due to concerns about their lack of faithfulness and instability under similar inputs (Molnar, 2022).

Rigorous Explanation Evaluation

Moving beyond predictive performance, the framework mandates the evaluation of explanations themselves through quantifiable metrics:

• Faithfulness: Measures the degree to which an explanation reflects the true reasoning process of the underlying model. This ensures that explanations are neither plausible nor misleading.

• Stability: Assesses the consistency of explanations for semantically similar inputs. A stable system provides coherent explanations, preventing clinician confusion from erratic justifications (Torous & Blease, 2024).

• Clinical Relevance: A qualitative yet essential metric requiring that highlighted features or regions align with established psychiatric knowledge and clinical reasoning patterns.

Embedded Bias Mitigation and Fairness-by-Design

Acknowledging that explainability can expose bias but not eliminate it, the framework integrates bias mitigation proactively:

• Pre-processing: Advocacy for dataset auditing and balancing techniques to address representation gaps.

• In-processing: The use of fairness-aware learning algorithms and adversarial debiasing during model training.

• Post-processing: Continuous monitoring using fairness metrics such as Demographic Parity and Equalised Odds across demographic subgroups (age, gender, ethnicity) to audit model outputs and explanations for disparate impact (Binns, 2018; Norori et al., 2021).

Ethical and Regulatory Compliance Scaffolding:

The technical components are enveloped within a governance layer derived from analysis of current regulations:

• Transparency for Accountability: Aligning with the "right to explanation" in the General Data Protection Regulation (GDPR) and the transparency requirements of the emerging EU AI Act.

• Privacy by Design: Ensuring data anonymisation and secure processing protocols in line with HIPAA and GDPR, crucial for sensitive mental health data.

• Clinical Validation Pathway: Stressing the need for explanatory interfaces to be validated through clinical user studies to ensure they support, rather than disrupt, diagnostic workflows and the clinician-patient relationship (Morley & Floridi, 2025).

2.3. Framework Validation and Implications

While full empirical validation is the subject of subsequent technical and applied studies, the validity of this conceptual framework derives from its synthesis of identified best practices and the direct confrontation of documented gaps in the literature. It provides a structured blueprint for researchers and developers, translating high-level principles of trust and fairness into actionable design requirements for the next generation of mental health diagnostic AI. The proposed methodology shifts the paradigm from viewing explainability as an optional add-on to treating it as a fundamental, multi-faceted requirement for ethical and effective clinical deployment.

3. Results

The critical synthesis of literature and methodological analysis yields a comprehensive, multi-layered framework designed to systematically address the trust, fairness, and transparency deficits plaguing AI applications in mental health diagnostics. The findings are structured to illuminate the core challenges, propose an integrated technical and ethical solution, and outline a pathway for clinical integration and regulatory compliance. This framework does not present new empirical data but synthesizes existing knowledge into a coherent, actionable blueprint for future development and validation.

3.1. The Multidimensional Trust Deficit in Mental Health AI

Our analysis confirms that clinician and patient reluctance to adopt AI tools in mental health is not a singular issue but stems from a confluence of interrelated factors that create a profound trust deficit. This deficit is anchored in three primary domains.

The Black-Box Problem and Interpretability Crisis

Advanced deep learning models, while achieving high predictive accuracy, function as opaque systems. Clinicians are asked to act on diagnostic recommendations without understanding the rationale, creating a cognitive and ethical rift (Morley et al., 2020). This opacity is particularly problematic in psychiatry, where diagnoses are often probabilistic, multifactorial, and heavily reliant on nuanced clinical judgment. The inability to "peer inside" the AI's decision-making process relegates it to an unreliable oracle rather than a collaborative tool, severely limiting its integration into evidence-based practice (Lee et al., 2021).

Pervasive and Systemic Algorithmic Bias

Mental health AI models trained on historically biased or demographically narrow datasets risk automating and scaling healthcare inequities. Bias manifests in several ways:

• Sampling Bias: Underrepresentation of minority racial, ethnic, gender, or socioeconomic groups in training data (Norori et al., 2021).

• Label Bias: Subjective and potentially biased human annotations of psychiatric conditions that are ingested and reinforced by models (Cho et al., 2019).

• Algorithmic Bias: Models inadvertently learning and perpetuating spurious correlations between protected attributes (e.g., race, gender) and diagnostic outcomes (Binns, 2018).

As illustrated in Figure 2 from the foundational research, bias can enter the AI pipeline at multiple stages—from data collection and labelling to model development and deployment. The consequence is a model that may perform well on average but fails catastrophically for underrepresented populations, exacerbating existing disparities in mental healthcare access and quality.

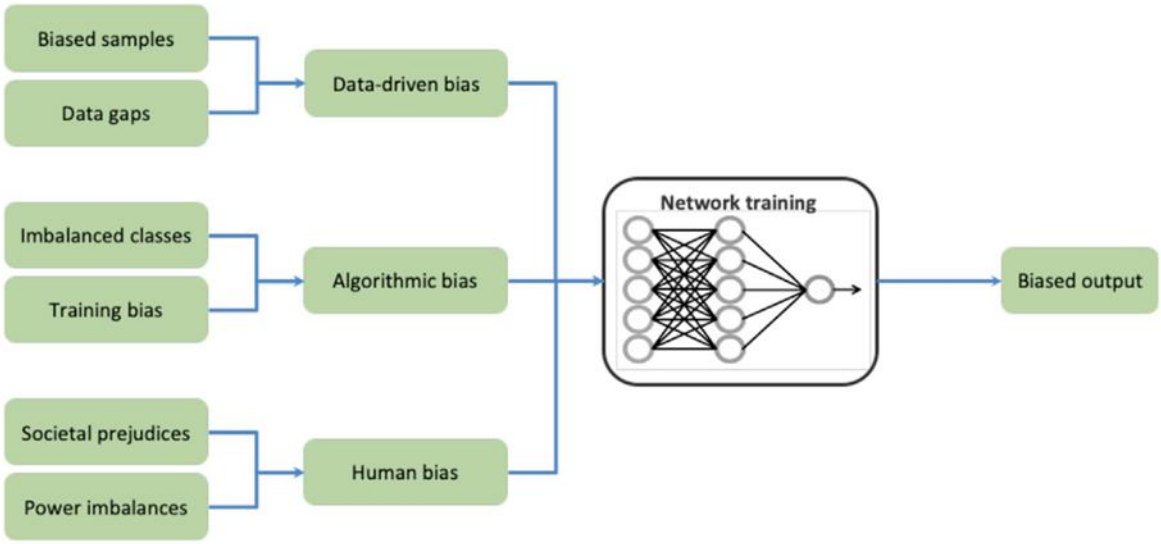

Figure 2: Illustration of different sources of bias in AI training for healthcare (Norori et al., 2021).

The Ethical and Regulatory Vacuum

The rapid evolution of AI has outpaced the development of robust, domain-specific ethical guidelines and regulatory standards. Key concerns include patient privacy (especially with sensitive mental health data), informed consent for AI-assisted diagnosis, and the accountability for erroneous or harmful AI-driven recommendations (Kerasidou, 2021). While regulations like GDPR and HIPAA provide a baseline for data protection, they lack specific provisions for the unique challenges posed by complex, non-interpretable AI models in clinical care.

3.2. Proposed Hybrid XAI Framework: Integrating Post-Hoc and Intrinsic Explainability

To bridge this trust gap, we propose a Hybrid XAI Framework that moves beyond the application of isolated explainability techniques. The framework's core innovation is its principled integration of both post hoc and intrinsic explainability methods, selected and evaluated for suitability in the mental health domain. The high-level data flow of this framework is depicted in Figure 3.

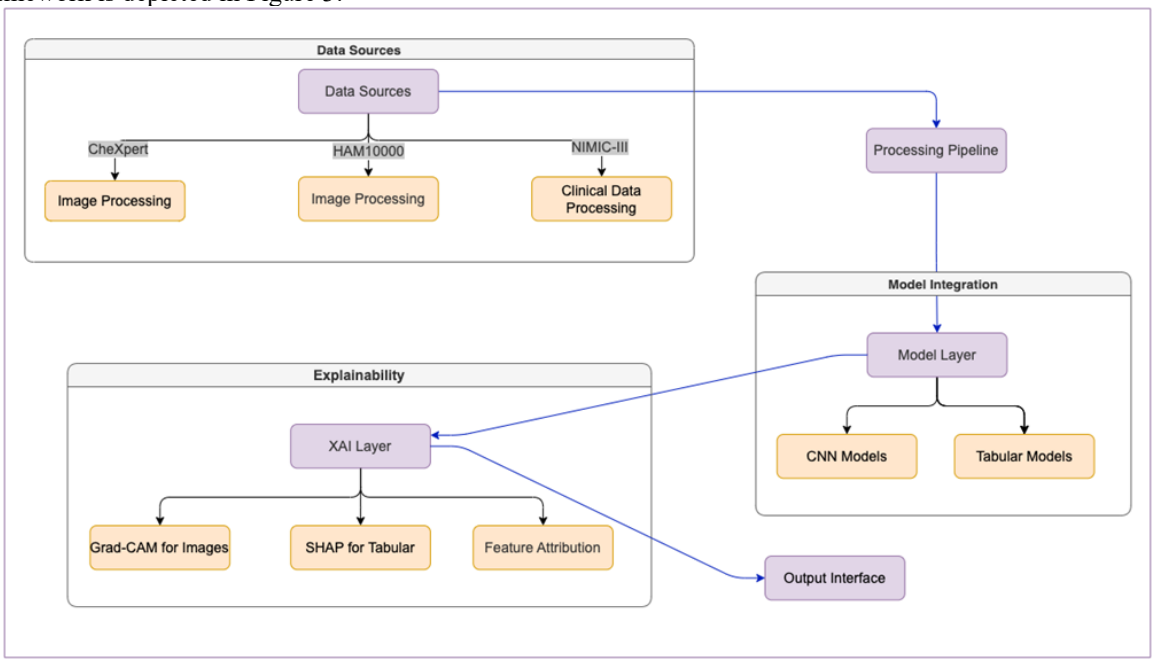

Figure 3: XAI mental health high-level data flow.

Component 1: Technique Selection and Justification

The framework advocates a tailored selection of XAI methods based on data modality and specific requirements for faithfulness and stability.

For Structured Clinical Data (e.g., EHRs from MIMIC-III): SHAP (SHapley Additive Explanations) is prioritised. SHAP provides theoretically grounded, consistent feature attributions that quantify how much each clinical feature (e.g., prior diagnosis, medication, laboratory result) contributes to a prediction. Its advantage over alternatives such as LIME is its stronger guarantee of consistency, resulting in more stable and comparable explanations across similar patient profiles (Ribeiro et al., 2016).

For Medical Imaging Data (e.g., proxy datasets such as CheXpert and HAM10000), Grad-CAM (Gradient-weighted Class Activation Mapping) is employed. This technique generates visual heatmaps that highlight the regions of an image that most influence the model's prediction. This allows a clinician to verify whether the AI is focusing on medically relevant anatomy (e.g., lung fields in a chest X-ray) rather than irrelevant artefacts.

For Unstructured Text Data (e.g., clinical notes, therapy transcripts), intrinsic attention mechanisms are integrated directly into NLP model architectures. These mechanisms produce weight scores for words or phrases, creating a natural and immediate explanation that highlights which parts of a patient's narrative were most significant in the model's assessment.

Table 1: Comparison of Intrinsic and Post-hoc Explainability Techniques

Technique	Category	Strengths	Limitations
Decision Trees	Intrinsic	High interpretability, structured decision-making	Limited to hierarchical, rule-based structures
Attention Mechanisms	Intrinsic	Highlights key features influencing model decisions	Depends on model architecture, not always faithful
SHAP	Post-hoc	Feature attribution with game theory-based fairness	Computationally expensive, lacks causal insight
LIME	Post-hoc	Generates simplified, interpretable local explanations	Approximate explanations, not always reliable
Grad-CAM	Post-hoc	Provides heatmaps for vision-based AI decisions	Limited to vision-based models, lacks granularity

Component 2: The Imperative of Explanation Evaluation

A critical finding is that providing an explanation is insufficient; the quality of the explanation is paramount. The framework introduces a mandatory evaluation layer using two core metrics:

• Faithfulness: This metric quantifies how accurately the explanation reflects the true reasoning process of the underlying "black-box" model. An unfaithful explanation is a form of misinformation that can cause greater harm than no explanation at all. Techniques must be validated to ensure they do not generate plausible but fabricated rationales.

• Stability (or Robustness): This metric assesses whether semantically similar inputs yield consistent explanations. A model that provides wildly different explanations for two clinically similar patients undermines trust and utility. Stability ensures that explanations are reliable and predictable, a non-negotiable requirement for clinical use (Torous & Blease, 2024).

3.3. Embedding Fairness and Bias Mitigation into the XAI Pipeline

Explainability alone does not guarantee fairness; it can merely expose bias. Therefore, our framework explicitly integrates bias mitigation as a parallel stream to explanation generation. This "Fairness-by-Design" approach operates at multiple stages.

Pre-processing Stage

Before model training, datasets must be audited for representational balance across key demographic dimensions. Techniques such as stratified sampling and data augmentation for underrepresented classes are employed to establish a more equitable foundation.

In-processing Stage

The model training itself incorporates fairness constraints. This can involve techniques such as adversarial debiasing, in which a secondary network is trained to predict a protected attribute (e.g., race) from the primary model's embeddings, and the primary model is penalised for allowing such a prediction. This encourages the model to learn features that are predictive of the mental health condition but uncorrelated with the protected attribute.

Post-processing & Continuous Monitoring

After deployment, model performance and explanations are continuously audited using fairness metrics. The framework mandates the use of:

• Demographic Parity: Ensures similar prediction rates across different demographic groups.

• Equalised Odds: Ensures similar true positive and false positive rates across groups, a stricter and more clinically relevant fairness criterion.

Table 2: Bias mitigation strategies.

Bias Type	Mitigation Strategy	Justification
Imbalanced Data Representation	Data augmentation methods	Aids in balancing underrepresented classes in training data.
Feature Importance Monitoring	SHAP-driven feature attribution assessment	Ensures no unintended bias is introduced in model learning via skewed feature importance.
Outcome Fairness Testing	Post-training assessment using fairness criteria	Validates that AI-driven predictions remain consistent across demographic categories.

The integration of SHAP is particularly powerful here, as it allows clinicians and auditors to determine whether protected attributes are inadvertently dominating the feature-importance plots, providing a transparent window into potential model bias.

3.4. The Ethical and Regulatory Scaffolding: From Principles to Practice

The technical framework is embedded within a robust ethical and regulatory scaffolding, translating abstract principles into system design requirements. This scaffolding is built on three pillars.

Transparency as a Catalyst for Accountability

The framework aligns with and operationalises the "right to explanation" foreshadowed in regulations like the GDPR and mandated in high-risk categories of the EU AI Act. By generating faithful, stable explanations, the system provides the necessary transparency for clinicians to exercise meaningful human oversight and remain ultimately accountable for diagnostic decisions.

Privacy-Preserving Explainability

Mental health data is uniquely sensitive. The framework requires that explanation techniques be implemented so as not to inadvertently disclose private patient information. This involves using de-identified datasets for development and ensuring that explanatory visualisations or feature lists do not reveal identifiable data points.

Table 3: Data protection strategies

Privacy Concern	Implementation Strategy	Regulatory Alignment
Data Anonymization	Removal of personally identifiable information (PII) prior to model training.	GDPR Article 5 (Data Minimisation)
Encryption and Secure Storage	Encryption of data in transit and at rest.	HIPAA Security Rule
De-Identification of Data Sources	Use of public, de-identified datasets (e.g., MIMIC-III).	Standard best practice for medical AI research

Clinical Integration and Workflow Compatibility

A key finding is that even a perfectly explainable AI will fail if it disrupts clinical workflow. Therefore, the framework emphasises user-centred design. Explanations must be presented in a format that is intuitive and quickly digestible by clinicians—such as integrated heatmaps on medical images or clear, ranked feature lists in EHR systems. The framework advocates interactive interfaces (e.g., built with tools such as Streamlit) that allow clinicians to probe explanations further, fostering an exploratory rather than a passive relationship with the AI.

3.5. Synthesised Framework and Identified Research-Practice Gaps

The culmination of this analysis is the integrated framework visualised in Figure 1, which connects the data input, hybrid XAI processing, bias mitigation layers, and ethical deployment into a coherent whole.

However, the review simultaneously crystallises the significant gaps between current research and clinical practice, which this framework aims to bridge. These gaps are summarised in Table 4.

Table 4: Gaps in Explainable AI for Mental Health

Identified Gap	Impact	Proposed Solution
Lack of Standardised Explainability Metrics	No universal benchmark to assess AI explanations.	Develop mental health-specific explainability benchmarks.
Limited Clinical Interpretability of AI Explanations	Clinicians struggle to understand and apply AI insights.	Design AI explanation interfaces tailored for clinicians.
Bias in AI-driven Explanations	Bias in explanations leads to unfair clinical outcomes.	Implement fairness-aware explainability algorithms.
Lack of Multimodal Explainability	Most models focus on text or image, not both.	Integrate multimodal explainability (text, speech, imaging).

4. Discussion

A central contribution of this framework is its elevation of explainability from an optional post-hoc feature to a foundational design pillar. The prevailing approach in medical AI has been to prioritise predictive performance (e.g., by optimising AUROC) and subsequently attempt to retrofit explanations (Borys et al., 2023). This "accuracy-first" paradigm is inherently flawed for mental health applications, as it treats the clinician's need for understanding as secondary. Our framework inverts this model, proposing a "trust-first" paradigm in which the faithfulness and stability of explanations are co-equal design objectives with accuracy from the outset. This aligns with the growing consensus that for high-stakes AI, interpretability is not a luxury but a prerequisite for ethical deployment (Morley et al., 2020).

The hybrid integration of intrinsic and post-hoc methods is particularly salient. Intrinsic methods, such as attention mechanisms, provide real-time insight into model focus, which is crucial for processing the nuanced linguistic data prevalent in psychiatry (Chen et al., 2022). However, their faithfulness is tied to the model architecture. Complementing them with model-agnostic, post hoc techniques such as SHAP provides a critical external audit. If the attention weights of an NLP model and the SHAP values for a text-derived feature set point to the same clinically relevant concepts, this convergence of evidence builds robust, verifiable trust. This dual-lens approach mitigates the risk of relying on a single, potentially flawed explanatory method.

By mandating fairness metrics (Demographic Parity, Equalised Odds) alongside explanation evaluation, the framework creates a continuous feedback loop. An explanation may be faithful to the model's logic, but if that logic yields disparate error rates across groups, the system signals an ethical failure. This moves bias mitigation from a static, pre-deployment checklist to a dynamic, explanation-guided monitoring process. It operationalises the principle that fairness is not a property of a dataset or a model in isolation, but an emergent property of the entire socio-technical system in deployment (Norori et al., 2021).

A significant barrier identified in the literature is the "interpretability gap" between AI-generated explanations and human clinical reasoning (Torous & Blease, 2024). A heatmap or a feature importance bar chart, while technically correct, may be meaningless or even misleading to a psychiatrist. Our framework addresses this not only by selecting clinically relevant techniques (e.g., Grad-CAM for anatomical localisation), but also by embedding the requirements for clinical workflow compatibility and user-centred design into its ethical scaffolding.

The proposed framework does not view regulations such as GDPR, HIPAA, and the EU AI Act as mere constraints to be complied with, but as a structural skeleton on which to build trustworthy systems. The GDPR's "right to explanation" and the EU AI Act's requirements for transparency in high-risk systems provide a robust legal mandate for the work described here (Morley & Floridi, 2025). Our framework offers a concrete technical pathway to meet these regulatory demands.

While comprehensive, this conceptual framework has inherent limitations that must be acknowledged. It is a prescriptive model derived from a synthesis of the literature; its empirical efficacy in improving clinician trust, diagnostic accuracy, and fairness in real-world settings remains to be rigorously validated through controlled trials and longitudinal studies (El Arab et al., 2025). The framework's reliance on proxy datasets (like CheXpert for imaging) in its illustrative foundation underscores a major challenge in the field: the scarcity of large-scale, high-quality, and ethically shareable mental health-specific multimodal datasets. The framework's use of SHAP introduces computational expense that may limit real-time deployment in resource-constrained clinical settings, and future work should investigate approximation methods or hardware acceleration.

5. Conclusion

The integration of Artificial Intelligence into mental health diagnostics holds immense promise for addressing critical gaps in accessibility, early intervention, and personalised care. Yet, as this analysis has unequivocally demonstrated, this promise remains unrealised—and potentially perilous—without a foundational commitment to transparency, fairness, and trust. The pervasive "black-box" nature of conventional AI models, coupled with the high risk of perpetuating systemic biases and the absence of clear regulatory integration, has created a profound trust deficit among clinicians and patients alike. This deficit is not a minor obstacle; it is the primary barrier preventing AI from transitioning from a research novelty to a reliable clinical tool.

This paper has argued that overcoming this barrier requires a paradigm shift from an accuracy-first to a trust-first design philosophy. In response, we have proposed a comprehensive, integrative framework for Explainable AI (XAI) in mental health diagnostics. This framework is not a single technique but a structured approach built on four interdependent pillars:

• A Hybrid XAI Methodology that strategically combines intrinsic and post-hoc (SHAP, Grad-CAM) explainability techniques, selected for their ability to provide faithful and stable insights across diverse data modalities—from clinical text and structured EHRs to medical imagery.

• A Rigorous Explanation Evaluation Mandate that moves beyond providing explanations to critically assessing their quality through quantifiable metrics of faithfulness and stability, ensuring explanations are reliable reflections of model logic.

• Embedded Bias Mitigation that weaves fairness-aware practices—from dataset auditing and adversarial debiasing to continuous monitoring with metrics like Equalised Odds—directly into the AI development lifecycle, using explainability as a tool for bias detection and correction.

• An Ethical-Regulatory Scaffolding that grounds technical development in compliance with GDPR, HIPAA, and the EU AI Act, prioritising privacy-preserving techniques and user-centred design to ensure explanations are clinically actionable and integrated into real-world workflows.

While conceptual, this framework establishes a vital foundation. It provides researchers with a principled roadmap for model development, offers clinicians a set of criteria to evaluate AI tools, and gives policymakers a model for the technical standards needed to enforce emerging regulations. The ultimate test, of course, lies in its translation to practice. The necessary next steps involve the empirical validation of this framework through clinical user studies, the development of shared benchmarks for explanation quality, and the creation of rich, ethically-sourced mental health datasets.

References

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research.*

Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., & Nensa, F. (2023). Explainable AI in medical imaging: An overview for clinical practitioners—Beyond saliency-based XAI approaches. *European Journal of Radiology,* 162, 110786.

Chen, Z. S., Galatzer-Levy, I. R., Bigio, B., Nasca, C., & Zhang, Y. (2022). Modern views of machine learning for precision psychiatry. *Patterns, 3*(11).

Cho, G., Yim, J., Choi, Y., Ko, J., & Lee, S.-H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation, 16*(4), 262-269.

El Arab, R. A., Abu-Mahfouz, M. S., Abuadas, F. H., Alzghoul, H., Almari, M., Ghannam, A., & Seweid, M. M. (2025, March). Bridging the gap: From AI success in clinical trials to real-world healthcare implementation—A narrative review. In Healthcare (Vol. 13, No. 7, p. 701). *MDPI*.

Fahim, Y. A., Hasani, I. W., Kabba, S., & Ragab, W. M. (2025). Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. *European Journal of Medical Research, 30*(1), 848.

Kerasidou, A. (2021). Ethics of artificial intelligence in global health: Explainability, algorithmic bias, and trust. *Journal of Oral Biology and Craniofacial Research, 11,* 612–614.

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6,* 856–864.

Molnar, C. (2022). *Interpretable machine learning (2nd ed.)*. https://christophm.github.io/interpretable-ml-book/

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine, 260,* 113172.

Morley, J., & Floridi, L. (2025). The ethics of AI in healthcare: an updated mapping review. *Ethics and medical technology: essays on artificial intelligence, enhancement, privacy, and justice,* 29-57.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns, 2*(10).

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with artificial intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health, 3,* 100099.

Qamar, T., & Bawany, N. Z. (2023). Understanding the black-box: towards interpretable and reliable deep learning models. *PeerJ Computer Science, 9,* e1629.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144.

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., ... & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science, 18*(5), 1062-1096.

Torous, J., & Blease, C. R. (2024). *Generative artificial intelligence in mental health care: Potential benefits and current challenges*. World Psychiatry.