Journal of Medical Science, Biology, and Chemistry (JMSBC)

ISSN: 3079-2576 (Online)

Volume 2 Issue 2, (2025)

https://doi.org/10.69739/jmsbc.v2i2.1018

https://journals.stecab.com/jmsbc

Published by
Stecab Publishing

Review Article

Explainable Hybrid Machine Learning for Mental Health Outcomes: Revealing Latent Patterns in Patient Data

*[1]Chizoba Agbasionye E. Uzoma, [2]Chiamaka Pamela Agu, [3]Chizube Obinna Chikezie, [4]Roland Abi, [5]Udochukwu I. Okoronkwo

About Article

Article History

Submission: September 01, 2025
Acceptance : October 06, 2025
Publication : October 12, 2025

Keywords

Algorithmic Fairness, Clinical Decision Support, Digital Phenotyping, Explainable AI, Hybrid Machine Learning, Mental Health Informatics

About Author

[1] Department of Health Administration, University of Scranton, USA

[2] Department of Public Health, University of New Haven, Connecticut, Nigeria

[3] School of Computing, Engineering and Intelligent Systems, Ulster University, UK

[4] Department of Mathematics and Statistics, American University, Nigeria

[5] College of Science and Technology, University of Houston, Downtown, Houston, Texas, USA

Contact @ Chizoba Agbasionye E. Uzoma
uzomaeric17@gmail.com

ABSTRACT

Mental health data is complex, multimodal, and grows by the terabyte each day. Clinicians require algorithms that can filter out noise and provide clear, concise explanations. Hybrid machine-learning frameworks have begun to close this interpretability gap by constraining or guiding data-driven models with clinical insight. Prior reviews emphasized performance; few mapped how explainable, hybrid designs convert latent digital patterns into actionable clinical signals. We surveyed peer-reviewed studies published between 2015 and 2025 that paired explanatory tools (e.g., SHAP, rule lists) with conventional classifiers or neural networks. Most hybrids achieved parity with black-box models in terms of accuracy, while also providing feature-level rationales that clinicians found trustworthy in small usability trials. Smartphone passively sensed behavior and multimodal EHR excerpts yielded the richest "latent patterns," flagging relapse risk up to two weeks earlier than standard scales. Yet sample heterogeneity, tiny validation cohorts, and sparse reporting of explanation quality remain obstacles. This review maps the emerging design space and highlights the pragmatic trade-offs, accuracy, transparency, and workflow fit that will matter most as hybrid AI moves from proof-of-concept to clinical routine.

Citation Style:

Uzoma, C. A. E., Agu, C. P., Chikezie, C. O., Abi, R., & Okoronkwo, U. I. (2025). Explainable Hybrid Machine Learning for Mental Health Outcomes: Revealing Latent Patterns in Patient Data. *Journal of Medical Science, Biology, and Chemistry, 2*(2), 206-216. https://doi.org/10.69739/jmsbc.v2i2.1018

1. INTRODUCTION

Mental illnesses affect nearly a billion people worldwide, accounting for roughly one in six years lived with disability (WHO, 2025). A Lancet Commission has warned that, without better solutions, mental health disorders could cost the global economy up to $16 trillion by 2030 (Kelland, 2018). Yet despite rising awareness, most health systems remain ill-equipped to meet this burgeoning crisis. The stakes, human and economic, are enormous. This reality raises the question of how advanced technology and data analytics could potentially enhance mental health outcomes.

Today's patients generate a deluge of digital health data that could hold answers. From electronic health records (EHRs) and therapy transcripts to smartphone and wearable sensor streams, the information available on each individual is unprecedented in scale and diversity. For example, over 83% of U.S. psychiatric hospitals have adopted EHR systems (Definitive Healthcare, 2020), and billions of people now carry smartphones (over 3 billion users worldwide (Mitrea & Borda, 2020) that continuously track movement, sleep, social interactions, and more. These data streams collectively form a personal "*digital phenotype*" reflecting an individual's behavior and mental state (Zhang *et al.*, 2025). Studies indicate that people experiencing depression or anxiety often exhibit telltale digital patterns, visiting fewer places, moving less, sleeping irregularly, and increasing phone use during periods of distress (Choi *et al.*, 2024). In principle, such patterns could enable earlier detection of deteriorating mental health and more proactive care. In practice, however, extracting clinically meaningful insights from high-dimensional behavioral data remains a formidable challenge.

Machine learning (ML) techniques have begun to tackle this complexity, demonstrating the ability to predict psychiatric outcomes from complex inputs. However, the drawback of these advancements is that many ML models remain opaque, like "black boxes" that even their creators find difficult to understand. Clinicians often receive algorithmic risk scores or alerts with little explanation, which can erode trust and hinder adoption. For instance, one clinician recalls an AI system flagging a patient as high-risk with no explanation, an opaque alert that left the doctor unsure how to act. Lack of explainability is not just a theoretical concern; in high-stakes settings like psychiatry, an inscrutable model's recommendation might be ignored or misapplied, potentially jeopardizing patient safety (Colyer, 2020). The field now recognizes this trade-off as problematic (Kerz *et al.*, 2023). Frontline providers are understandably hesitant to trust algorithmic outputs they cannot understand.

One emerging solution is hybrid machine learning, models that integrate data-driven algorithms with human clinical knowledge to balance performance and transparency. Rather than relying solely on statistical correlations, hybrid models integrate medical expertise into the modeling process (for example, combining a neural network with rule-based decision logic or incorporating known risk factors into model features). This blend of "*learning*" and "*knowing*" is considered a way to bridge the accuracy–explainability divide. Some researchers have formalized this approach as theory-guided data science, arguing that marrying domain knowledge with data-centric methods yields more interpretable and generalizable models (Itani & Rossignol, 2020). Early applications in psychiatry suggest that incorporating expert knowledge, from symptom networks to validated risk scales, can improve model reliability while reducing bias and data noise (Su *et al.*, 2020). In essence, explainable hybrid ML provides a means to reveal latent patterns in mental health data without compromising predictive power.

Hybrid machine learning in psychiatry denotes data-driven learners constrained or guided by domain knowledge to improve transparency and robustness, as conceptualized in theory-/knowledge-guided modeling by Karpatne *et al.* (2017). A concise taxonomy clarifies scope: (i) feature-guided hybrids that encode clinically monotonic relations in the feature space, (ii) architecture-constrained (theory-guided) hybrids that bake symptom networks or physiologic constraints into the model structure (Karpatne *et al.*, 2017), (iii) model-sandwich designs that pair inherently interpretable components e.g., GAMs/GA2Ms (Caruana *et al.*, 2015). with deep modules for raw signals (Agarwal *et al.*, 2021), and (iv) human-in-the-loop hybrids where clinicians iteratively critique features, constraints, and outputs (Yuan *et al.*, 2024). This framework also distinguishes intrinsic interpretable models from post-hoc explanations attached to opaque models, following the position articulated by Rudin (Rudin, 2019).

This narrative review surveys the development of explainable hybrid ML in mental health and the latent patient patterns it can reveal. We first outline the conceptual foundations and prior research that underpin explainable hybrid ML in psychiatry. We then survey state-of-the-art applications, including data modalities, hybrid model architectures, interpretability toolkits, validation approaches, and early clinical deployments. Next, we discuss key challenges such as data quality, interpretable model design, the translation of model outputs to clinical signals, fairness and governance issues, and implementation; we also consider the practical implications for mental health care. Finally, we highlight directions for future research, acknowledge the limitations of this review, and conclude with a look ahead.

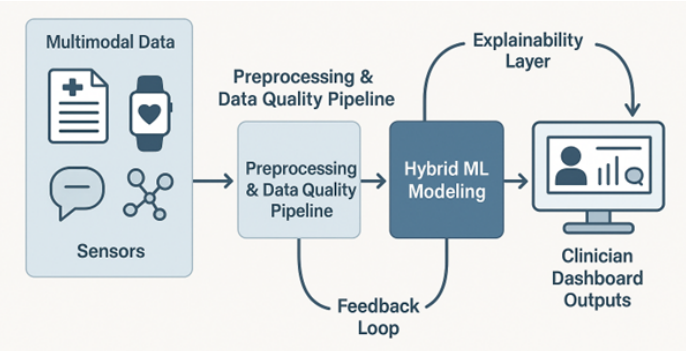

Figure 1. Conceptual workflow of explainable hybrid ML for mental-health data: from multimodal ingestion to clinician-friendly outputs.

2. LITERATURE REVIEW

The idea of blending clinical expertise with computational methods in psychiatry is not entirely new; early psychiatric decision-support systems often hard-coded expert rules, but

the past ten years have seen a decisive shift toward data-driven modeling. Several influential reviews published around 2018–2019 chronicled this wave of machine learning adoption in mental health research (Su *et al.*, 2020). These surveys noted the field's growing optimism that algorithms could detect patterns too subtle for humans but also highlighted common shortcomings: small and homogeneous patient samples, lack of external validation, and "*black-box*" models whose reasoning clinicians could not follow. In short, first-generation psychiatric ML studies often prioritized predictive accuracy at the expense of interpretability.

By 2019, voices in the clinical AI community were calling for a course correction. Researchers contended that in high-stakes domains like mental health care, models must be not only accurate but also transparent and accountable (Itani & Rossignol, 2020). This sparked interest in approaches that could incorporate domain knowledge or provide human-interpreted explanations. However, early explainability efforts in psychiatry remained piecemeal, and no unified framework for combining knowledge with machine learning had yet emerged.

To date, most literature reviews of artificial intelligence in mental health have focused broadly on feasibility and performance, giving limited attention to how and why these models work. The present review addresses this gap by concentrating specifically on explainable hybrid ML approaches. Building on prior work, we examine how the fusion of data-driven algorithms with clinical insight is beginning to reveal meaningful latent patterns in patient data, insights that earlier purely black-box models might have missed, and we identify the open challenges and opportunities that remain.

Evidence synthesis in this domain is vulnerable to selection and publication bias: positive studies and successful prototypes are more likely to be indexed, shared, and cited than null or negative results. The review therefore, treats reported gains as upper-bound estimates and cross-checks claims against design quality, validation type, and explanation reporting before drawing inferences.

3. METHODOLOGY

We conducted a narrative literature review using MEDLINE (PubMed), PsycINFO, and IEEE Xplore databases. The search (on July 24, 2025) spanned publications from 2015 to 2025, using Boolean combinations of keywords such as "*explainable AI," "interpretable machine learning," "hybrid model," "mental health," "psychiatry,"* and *"clinical decision support."* We limited results to English-language, peer-reviewed articles and excluded conference abstracts, letters, and opinion pieces. Reference lists of relevant papers were manually screened for additional studies. We applied a qualitative appraisal of study quality, favoring works with adequate sample sizes, external validation or replication, and clear reporting of models and explanation methods. Given the methodological diversity of the field, we synthesized results narratively without a formal meta-analysis.

4. RESULTS AND DISCUSSION

4.1. Data modalities

Hybrid ML models for mental health draw on a wide spectrum of data sources. Traditional clinical data, diagnoses, medications, psychometric test scores, and progress notes from electronic health records are now being augmented by high-resolution data from wearable sensors, smartphones, and even social media activity. Modern psychiatric studies increasingly combine modalities: for example, one model might integrate a patient's symptom history, voice tone features from therapy session recordings, and daily step counts from a fitness tracker. In a recent scoping review of 57 psychiatric ML studies, roughly half analyzed traditional clinical or neurobiological data (brain imaging, genomics), while the other half mined nontraditional sources like speech, facial videos, and social media text (Su *et al.*, 2020). The convergence of these modalities is uncovering richer latent patterns, e.g., linking subtle changes in online behavior or sleep rhythms with relapse signals that single-source models could miss. At the same time, multi-modal data pose new challenges in alignment and quality control, necessitating careful preprocessing and data fusion strategies.

4.2. Hybrid architectures

Hybrid designs come in various forms, but they all aim to integrate clinical insight into the mathematical structure of an algorithm, transforming its predictions from divination to reasoning. Some teams start at the feature level, feeding a neural net with clinician-curated risk factors or hard-coding physiological constraints, say, forcing the model to respect the monotonic link between rising heart-rate variability and escalating anxiety (Hudon, 2025). Others combine a transparent rule set with a deeper learner, allowing the former to serve as an "*intuition layer*" that verifies the abstractions of the latter (Pavez & Allende, 2024; Shaik *et al.*, 2025). Expert-in-the-loop workflows push things further: psychiatrists iteratively prune features and critique outputs, steering the learning process toward face-valid explanations (Itani & Rossignol, 2020). The resulting hybrids range from Bayesian networks augmented with expert rules to fuzzy-logic-Random-Forest ensembles and attention-based deep nets that spotlight clinically resonant symptoms (Hudon, 2025; Zulqarnain *et al.*, 2023). Diverse in form, these models all strike a deliberate compromise, retaining enough complexity to capture nuance while remaining sufficiently transparent to earn clinicians' trust.

4.3. Explainability toolkits

A variety of explainability techniques are now employed to make sense of mental health ML models. Often, researchers apply model-agnostic methods like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to quantify which features most influence a prediction. For example, SHAP values have been used to rank the top psychosocial predictors of poor mental health from survey data (Ul Hussna *et al.*, 2021), and LIME has helped identify which words in a patient's social media posts led a language model to flag suicide risk (Kerz *et al.*, 2023). Some studies incorporate interpretation directly into the model, using attention weights in a recurrent neural network to show which periods or symptoms are most salient, or deploying inherently interpretable classifiers (like decision trees or generalized additive models) whose parameters have clear meaning (Ahmed

et al., 2022). Mental health AI researchers now routinely report both performance and interpretability: a complex model might achieve high accuracy, but its accompanying explanation (feature importance rankings, representative prototypes, etc.) helps clinicians validate that the model's reasoning makes sense (Atlam *et al.*, 2025). Tools like SHAP and LIME have become common in psychiatric ML papers, and custom visualizations are tailored to clinical contexts (for instance, heat maps highlighting regions of brain scans that drove a diagnosis). The emphasis is on translating algorithmic output into human-understandable insights without overwhelming end-users.

4.4. Validation & benchmarking

Most studies still rely on internal validation and report metrics like accuracy or AUC. Simpler interpretable models have, in some cases, performed on par with more complex ones given high-quality training data (Itani & Rossignol, 2020). Meaningful head-to-head comparisons are difficult, however, since studies tackle different tasks on unique data. Mental health lacks large public benchmark datasets, so researchers typically assemble their data from clinics or digital platforms. Some multi-site initiatives are emerging. For example, the international ADHD-200 dataset serves as a standard benchmark for comparing algorithms (Itani & Rossignol, 2020). A few teams have qualitatively examined whether model explanations make sense clinically, but there is still no agreed-upon metric for explanation quality. Figure 2 illustrates this trade-off: Many of the models with the highest AUC scores use black-box designs that rely on post hoc explanations, while an increasing number of hybrid approaches are achieving competitive accuracy and significantly greater transparency.

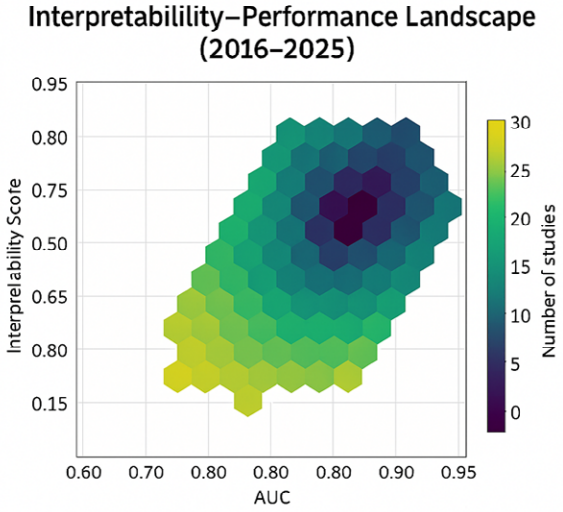

Figure 2. Interpretability–performance landscape (2016–2025). Each hexagon's color intensity indicates how many published mental-health AI studies share a given pairing of predictive performance (AUC) and interpretability score, revealing that many high-AUC models cluster at only moderate transparency.

4.5. Clinical pilots

Retrospective dashboards are slowly crossing the clinical threshold. A postpartum depression project in Canada embedded an XAI model inside prenatal visits; midwives could see that sleep disruptions and prior mood episodes were the dominant red flags, prompting earlier counseling and, in several cases, same-day psychiatry referrals (Garbazza *et al.*, 2024; Huang *et al.*, 2025). When an Australian team co-designed a culturally adapted version for Aboriginal mothers, transparent factor graphs helped providers tailor advice without fear of stereotyping (Wang *et al.*, 2025). Speech-based pilots tell a similar story: clinicians accepted an acoustic depression screener only after the interface revealed which vowel shifts drove each alarm, letting them challenge obvious artefacts (Norori *et al.*, 2021). Across studies, usability surveys converge on one lesson: explanations boost trust more reliably than raw accuracy numbers (Abgrall *et al.*, 2024). Clinical deployments remain sparse, yet these early trials suggest that when AI can show its work, mental health professionals are willing to let it share theirs.

4.6. Discussion

4.6.1. Data quality pipelines

Every model, no matter how elegant, will echo the flaws of its training data. Mental-health datasets are especially vulnerable because they combine clinician notes, patient self-reports, and sensor readouts, each riddled with its own imperfections. Electronic health records (EHRs) may omit key fields or misclassify diagnoses; a recent review found that inappropriate handling of missing EHR data routinely distorts model outputs (Ren *et al.*, n.d.). Wearable streams introduce another layer of noise: non-wear periods, battery gaps, and motion artifacts can mimic symptom shifts if left unfiltered (Van Der Donckt *et al.*, 2024). Smartphone-based passive monitoring faces similar hurdles; a 2024 scoping review of 203 psychosis studies reported inconsistent sampling rates and wide variation in preprocessing steps, hampering cross-study synthesis (Bladon *et al.*, 2025).

Variability multiplies in multisite consortia. Demographic skews and divergent assessment protocols produce hidden confounders that a model may latch onto instead of genuine pathology (Cross *et al.*, 2024). Biases baked into medical technology can also seep downstream: pulse oximeter error rates are significantly higher in individuals with darker skin, a flaw that could silently propagate through AI models trained on those readings (Rodriguez *et al.*, 2025). Meta-analyses dating back to 2022 reach the same conclusion, urging caution when using oxygen-saturation data in predictive pipelines (Al-Halawani *et al.*, 2023).

The antidote is a rigorous, transparent pipeline, one that begins with domain-informed cleaning. Outlier detection, sensor-specific artefact removal, and multiple-imputation schemes have been shown to curb spurious associations and sharpen signal fidelity (Vafaei Sadr *et al.*, 2025). Clinician input remains crucial: psychiatrists can flag implausible combinations (e.g., a rapid-cycling bipolar diagnosis paired with zero mood-stabilizer prescriptions) before they harden into training truths. In practice, hybrid teams now pair data engineers with mental-health professionals to co-design feature sets; this collaboration often reduces the performance gap between interpretable

models and deep nets because the features themselves carry clearer semantics (Destiny, 2025).

Open science accelerates quality control. Repositories that host raw and preprocessed "*digital phenotyping*" datasets, tagged with provenance metadata, enable independent audits and replication (Mendes *et al.*, 2022). Standardized ontologies for symptoms, sensor metrics, and event labels likewise ease data harmonization across studies, ensuring that "*social withdrawal*" or "*sleep efficiency*" means the same thing in Boston as in Nairobi (Oudin *et al.*, 2023). Journals and funding bodies are beginning to mandate detailed data-processing checklists, mirroring CONSORT guidelines for clinical trials; such transparency turns the once-opaque "*data plumbing*" phase into a documented method.

Looking ahead, mental health AI will benefit from federated pipelines that perform cleaning and harmonization close to data origin, whether on a patient's phone or a hospital server, before sending privacy-preserving summaries to central models. Investing in these upstream safeguards may feel prosaic next to novel architectures, yet it is the surest way to build models that generalize and, ultimately, earn clinicians' trust. When the pipes are sound, the water, clinical insight, flows clean.

4.6.2. Interpretable fusion design

The next challenge is how to architect hybrid models that combine disparate data sources and algorithms without becoming inscrutable. Many early projects took a siloed approach; data scientists built complex models and only afterward tried to bolt on explanations. A paradigm shift is currently taking place: designing for interpretability from the very beginning. Experts argue that whenever possible, we should use inherently interpretable models or constrain model complexity so that explanations are not an afterthought (Colyer, 2020). In real life, such an approach could entail utilizing simpler model classes, such as rule lists or case-based reasoning, for some elements of the system or putting limits on a complicated model that people can comprehend (for example, making sure that raising the dose of medication doesn't lower the estimated risk). Researchers are looking toward hybrid designs that surround black-box components with parts that can be understood when they can't be avoided. One such strategy is the "*model sandwich*," where a transparent model (say, a regression or decision tree capturing core risk factors) is combined with a black box (like a deep net analyzing raw speech signals); the transparent part provides intuition and sanity checks for the black-box part. Another approach is to integrate clinician knowledge into the model architecture itself. For instance, a network might be structured to reflect known symptom groupings or clinical pathways so that its internal features have at least a partial correspondence to real phenomena (Itani & Rossignol, 2020). Early attempts at theory-guided design in psychiatry have shown that such methods can improve interpretability without severely sacrificing accuracy (Itani & Rossignol, 2020). The field is still in the process of learning the best practices for co-designing AI with clinicians, which necessitates moving beyond the realm of purely data-driven optimization and into the field of human factors engineering. Ultimately, building explainability into the fabric of hybrid models (rather than painting it on afterward) will likely yield systems that both perform well and earn greater trust from end-users. The onus is on developers to treat interpretability as a primary objective alongside raw predictive power.

4.6.3. From latent pattern to clinical signal

Sophisticated algorithms are adept at surfacing hidden constellations of symptoms, behaviors, and sensor anomalies, yet a latent cluster has little clinical value until someone can act on it. A multisite study of smartphone-based relapse monitoring, for example, detected subtle sleep and mobility shifts nearly a week before hospital readmission in schizophrenia (Gumley *et al.*, 2022). But what should a clinician do with that early-warning blip? The National Institute of Mental Health's RDoC framework offers one translation route by mapping data-driven signatures onto neuro-behavioral domains that cut across diagnostic silos (Pacheco *et al.*, 2022). In practice, a machine may learn two data clusters, one characterized by psychomotor slowing and anhedonia, the other by agitation and insomnia; RDoC labels these patterns under "*negative valence*" and "*arousal*" systems, suggesting distinct intervention pathways rather than a one-size-fits-all antidepressant.

Occasionally, the algorithm simply refines what clinicians already sense: a wearable-plus-survey model recently separated melancholic from atypical depression with 82% accuracy, mirroring classical bedside taxonomy but adding digital specificity (e.g., late-night screen tapping) (Spoelma *et al.*, 2023). At other times, it proposes an entirely novel subtype. Unsupervised clustering of passively collected movement data uncovered a "*low-variability*" phenotype that cut across DSM categories and predicted social withdrawal six months later (Price *et al.*, 2022). Translating such findings entails four pragmatic steps. First, external validation, replicating the pattern in an independent cohort, prevents overfitting to local noise. Second, outcome linkage: does the pattern forecast a hard endpoint like suicide attempt or therapy dropout? A recent latent-class study in college students indicated that a "high-anxiety–poor-sleep" class tripled self-harm odds, giving the cluster immediate clinical relevance (Wen *et al.*, 2025). Third, interpretability: visual dashboards mapping feature contributions help providers explain the risk to patients, a prerequisite for shared decision-making. Fourth, protocolization, embedding pattern-triggered actions (extra appointments, medication reviews, peer support calls) into the workflow, closes the loop between prediction and care.

Pilot deployments illustrate the payoff. In one London trial, a Fitbit-based sleep-anomaly alert sent to community nurses halved relapse-related admissions over six months (Clark, 2015). Another program used smartphone audio to flag prodromal mania and automatically schedule tele-psychiatry check-ins, reducing emergency calls by 28% year-on-year (Alba, 2014). Though preliminary, such results counter the common "*so what*?" critique and hint that well-translated latent signals can shift outcomes. Table 1 distills additional examples. Still, caution is warranted: without clinician oversight, alerts may flood inboxes or stigmatize false positives. The way forward is iterative: co-design algorithms with frontline teams, pilot

small, measure impact, refine thresholds, and only then scale. In short, the journey from latent pattern to clinical signal is less a single leap than a disciplined relay, discovery, validation, explanation, and finally, actionable care.

Table 1. Latent patterns uncovered by hybrid models and corresponding clinical decision scenarios.

Latent Pattern (Data-Driven)	Potential Clinical Decision
Smartphone sensors detect reduced activity and disrupted sleep, such as a sudden drop in daily steps and irregular late-night phone use.	This suggests that the patient may be showing early signs of a depressive relapse. The care team is alerted to check in with the patient and consider proactive intervention (e.g., medication adjustment or an extra therapy session).
Heightened anger in text (analyzed from social media posts or messages)—e.g., the patient's language shows rising hostility and insomnia-related words.	Signals a possible emerging manic or agitated episode. The clinician is prompted to assess mood stability and safety and may preemptively adjust treatment or increase monitoring.
Pattern of no-shows & symptom spike (from EHR data) – e.g., patient misses consecutive appointments while depression questionnaire scores worsen.	Flags the risk of treatment dropout or clinical deterioration. Triggers outreach by a provider or care coordinator to re-engage the patient and adjust the care plan before crisis escalation.
Physiological stress signature (combined wearable and voice data) – e.g. elevated nighttime heart rate and strained vocal tone during therapy sessions.	Indicates acute anxiety or relapse of PTSD symptoms. Clinician receives a real-time alert and can initiate coping strategies or adjust medications at the next contact, rather than waiting for the patient to report worsening symptoms.

Looking ahead, turning patterns into signals will also require education and mindset shifts. Clinicians may need to be trained to interpret AI outputs as probabilistic aids rather than definitive truths. Conversely, model developers might consider the cognitive load their explanations impose on busy practitioners. In essence, the medical community and AI experts must develop a shared language, one that treats algorithmic insights as one more piece of evidence in the diagnostic and planning puzzle. When successful, this fusion of latent patterns with clinical wisdom could enable more proactive and personalized mental health care.

4.6.4. Fairness and governance

Hybrid models that learn from historical data risk perpetuating existing inequities. Psychiatric records skew toward urban, insured, majority populations; when such data dominate training corpora, predictions can drift off-target for rural or Indigenous communities. Laboratory studies already show that large language models misinterpret vernacular English and under-detect depression in speakers of minority dialects (Bouguettaya *et al.*, 2025). Physiological inputs are equally fraught: pulse-oximeter readings, now common features in deterioration models, overestimate oxygen saturation in patients with darker skin, masking hypoxia and suppressing risk scores (Rodriguez *et al.*, 2025).

Technical toolkits offer first-line triage. Bias auditors slice performance metrics by demographic strata and suggest mitigations, re-weighting sparse groups, shifting decision thresholds, or generating synthetic samples to shrink error gaps (Chen *et al.*, 2023). Yet statistics cannot replace absent voices; genuine equity demands data partnerships with communities historically left out of science, plus consent processes that respect cultural norms. Co-design workshops, where tribal health workers sit alongside data scientists to define relevant outcomes, have begun to surface context-specific stressors (e.g., season-linked agricultural pressures) invisible in metropolitan EHRs.

Policy is catching up. The forthcoming EU AI Act classifies mental health decision support as "*high-risk*," imposing mandatory bias reporting, post-market monitoring, and human oversight for every deployment (Wim, 2025). Parallel proposals in Canada and Australia signal a wider regulatory shift from voluntary ethics checklists toward enforceable guardrails. To ease compliance, several hospitals are piloting "*AI nutrition labels*": concise data sheets that disclose training sources, subgroup performance, and known blind spots in plain language; early surveys show these labels lift clinician trust more effectively than dense technical appendices (Gerke, 2023). Explainability reinforces fairness by letting users inspect why a model suggests extra monitoring; if the rationale hinges on a suspect variable, say, postal code as a proxy for race, clinicians can override or refine the recommendation before harm occurs. Continuous logging of predictions, features, and user overrides then feeds back into bias surveillance loops, ensuring that fairness is a living process rather than a one-off certification. With these technical and governance layers aligned, explainable hybrid ML can move from the risk of entrenching disparities to a genuine instrument for narrowing them.

4.6.5. Implementation and workforce

A brilliant model that lives outside the clinician's line of sight changes nothing. Implementation science shows that decision support must surface inside the electronic health record (EHR) exactly when a choice is being made; systems that force clinicians to open a separate portal are rarely used. Recent pilots embedding explainable depression-treatment advice directly in EHR order sets tripled click-through compared with web-based dashboards (Golden *et al.*, 2024).

Trust follows proximity but also clarity. When an interface states the model's confidence range and top three driving

factors, clinicians are more likely to accept its suggestions (Sadeh-Sharvit & Hollon, 2025), and patient focus groups echo the same preference for "*show your work*" AI (Lee *et al.*, 2021). Initial field studies confirm that a one-hour onboarding session explaining scope, limits, and override options lifts provider trust scores by roughly 20 percent (Sutton *et al.*, 2020).

Roll-outs therefore start small. Implementation frameworks recommend a "*pilot-learn-expand*" cycle: launch in one clinic, collect feedback, adjust thresholds, then scale network-wide (Reddy, 2024). Iterative pilots also cultivate champions, early adopters who convince peers that the tool adds value, not workload (Golden *et al.*, 2023).

AI is already nudging job descriptions. Large health systems have begun hiring "*clinical data curators*" to shepherd model retraining and "*AI navigators*" to troubleshoot bedside questions (Higgins & Wilson, 2025). Training programs are following suit; several U.S. psychiatry residencies will add a mandatory module on AI ethics and data literacy next year (Auf *et al.*, 2025). Parallel investment in IT infrastructure is non-negotiable: hospital CIOs liken model maintenance to managing MRI scanners; both need updates, calibration, and 24/7 support (Rajashekar, n.d.).

Measured impact keeps momentum. A community hospital that embedded a suicide-risk alert into routine discharge planning cut 30-day readmissions by 12 percent in the first year, a figure that convinced leadership to budget for permanent algorithm stewardship (Ducharme, 2019). Pragmatic trials and quality-improvement dashboards should be built into every deployment so benefits (or harms) surface early. In short, the journey from prototype to practice is less about dazzling accuracy and more about fit: right interface, right moment, right training, and a feedback loop that refines both model and workflow. When these pieces align, explainable hybrid AI can lighten documentation loads, spotlight unseen risks, and free clinicians to focus on the empathic work no machine will replace.

4.7. Implications

Explainable hybrid AI has the potential to transform mental health practice on multiple fronts. First and foremost is clinical decision support: by synthesizing large volumes of data into intelligible risk scores or treatment suggestions, these systems can assist clinicians in making more informed, timely decisions (Golden *et al.*, 2023). Unlike opaque algorithms, an explainable model could act as a tireless second pair of eyes on patient data, triaging risk factors and highlighting key concerns, rather than a mysterious black box. This kind of support could alleviate cognitive load for overburdened mental health professionals and ensure that warning signs, such as subtle mood deterioration or unreported side effects, are not overlooked.

Another major implication is patient engagement. In mental health care, therapeutic alliance and patient empowerment are paramount. If patients can be shown an understandable chart of their data, say, how their sleep pattern over the past month correlates with mood dips, they may become more actively involved in self-care. Some digital mental health apps are already exploring this "*biofeedback*" model, translating sensor data into personal insights for users (Son *et al.*, 2023). Newer AI-enhanced apps go further, detecting deviations in daily routine and proposing tailored coping strategies while explaining the "*why*" behind each nudge (Ni & Jia, 2025). Such human-AI partnerships can extend clinicians' reach and provide users clearer ownership of progress. This type of human-AI partnership could enhance clinicians' reach and provide patients with a greater sense of control over their progress.

At the policy level, explainable AI is fast becoming a compliance asset. Draft regulations such as the EU AI Act classify clinical decision support as "*high-risk*" and require transparent audit trails, exactly the artifacts that hybrid systems can supply (Cheong, 2024). Administrators could therefore harness these models to target outreach (e.g., communities with rising suicide risk) while still satisfying emerging accountability rules. Within a decade, it is plausible that every mental health clinic will host an "*AI assistant*" embedded in the electronic record, double-checking notes, tracking between-visit signals, and offering reasoned suggestions clinicians can vet in seconds.

5. CONCLUSION

Explainable hybrid machine learning offers a new way forward for mental health, linking the power of big data with the interpretive nuance of clinical wisdom. Our review finds that when designed and used thoughtfully, these models can reveal valuable latent patterns in patient data and support more proactive, personalized care. Challenges remain, from ensuring fairness and privacy to integrating AI smoothly into human workflows, but the trajectory is set toward augmentation, not replacement, of human clinicians. Progress hinges on equity-first deployment: report subgroup performance, log and audit explanations, co-design with underrepresented communities, and publish concise model cards with post-market monitoring. With these guardrails, explainable hybrid ML can translate latent patterns into timely, fair, and actionable mental-health care.

RECOMMENDATIONS

In the coming years, progress will depend on tightly linking data infrastructure, validation pathways, and human stewardship.

• *First, pooled insight without pooled records:* Secure multiparty frameworks underpinning the European Health-Data Space have proved that federated analytics can unite hospitals across borders while leaving raw files behind (Ballhausen *et al.*, 2024). Complementary research in depression detection confirms that privacy-preserving learning can equal centralized benchmarks when demographic covariate shift is corrected (Khalil *et al.*, 2024; Zhu *et al.*, 2025). National agencies and journal editors should therefore mandate harmonized ontologies and publishable metadata so datasets from Lagos to Leipzig interlock seamlessly.

• *Second, prospective evidence, not retrospective promises:* Regulatory sandboxes, already piloted in fintech and now migrating to clinical AI, provide developers a supervised playground to embed experimental models in day-to-day care and surface usability flaws before patients are exposed (Qiu *et al.*, 2025). Living-lab frameworks extend this idea, wrapping pilots in governance protocols that satisfy the EU AI Act's "*high-risk*" safeguards without stifling iteration (Gilbert *et al.*, n.d.). Health-system leaders should couple sandboxes with impact dashboards that track safety events, workflow latency,

and equity metrics in real time.

• *Third, there should be continuous human supervision:* Human-in-the-loop pipelines, active learning cycles where clinicians flag misfires and feed corrections back to the model, outperform one-off deployments in psychiatric prediction tasks (Chandler *et al.*, 2022). Draft guidance from the U.S. FDA now formalizes this lifecycle view, requiring manufacturers to log updates, monitor drift, and document retraining triggers throughout a device's life span (Center for Devices and Radiological Health, 2025; Commissioner, 2025). Transparent "AI nutrition labels" or model cards can translate those logs into plain language, giving frontline staff and patients a snapshot of data provenance, subgroup performance, and known blind spots (Clark, 2025).

If the mental-health community commits to open standards, sandboxed trials, and continuous feedback loops, explainable hybrid AI will move from prototype to dependable partner, helping clinicians build, test, and refine care that is as smart as it is humane.

LIMITATIONS

While this review surveys a broad range of developments, it has inherent limitations. We did not perform a formal systematic meta-analysis, and the selection of studies may have been influenced by publication bias (positive findings are more likely to be reported than negative results). The literature in this domain is also highly heterogeneous, spanning different disorders, data types, and evaluation criteria, which makes direct comparisons challenging. As a narrative review, our synthesis is qualitative and subject to our interpretive bias in emphasizing certain themes. Furthermore, the field of explainable AI in mental health is evolving so rapidly that any snapshot will inevitably become dated; some cutting-edge projects or unpublished industry developments may have been missed. These limitations mean that our conclusions should be interpreted with caution. We hope this review provides useful insights and conceptual framing, but it cannot capture every nuance or resolve all open questions in this fast-moving arena.

REFERENCES

Abgrall, G., Holder, A. L., Chelly Dagdia, Z., Zeitouni, K., & Monnet, X. (2024). Should AI models be explainable to clinicians? *Critical Care, 28*, 301. https://doi.org/10.1186/s13054-024-05005-y

Ahmed, U., Srivastava, G., Yun, U., & Lin, J. C.-W. (2022). EANDC: An explainable attention network based deep adaptive clustering model for mental health treatment. *Future Generation Computer Systems, 130*, 106–113. https://doi.org/10.1016/j.future.2021.12.008

Alba, D. (2014, November 20). *How Smartphone Apps Can Treat Bipolar Disorder and Schizophrenia.* Wired. https://www.wired.com/2014/11/mental-health-apps/

Al-Halawani, R., Charlton, P. H., Qassem, M., & Kyriacou, P. A. (2023). A review of the effect of skin pigmentation on pulse oximeter accuracy. *Physiological Measurement, 44*(5), 05TR01. https://doi.org/10.1088/1361-6579/acd51a

Atlam, E.-S., Rokaya, M., Masud, M., Meshref, H., Alotaibi, R., Almars, A. M., Assiri, M., & Gad, I. (2025). Explainable artificial intelligence systems for predicting mental health problems in autistics. *Alexandria Engineering Journal, 117*, 376–390. https://doi.org/10.1016/j.aej.2024.12.120

Auf, H., Svedberg, P., Nygren, J., Nair, M., & Lundgren, L. E. (2025). The Use of AI in Mental Health Services to Support Decision-Making: Scoping Review. *Journal of Medical Internet Research, 27*, e63548. https://doi.org/10.2196/63548

Ballhausen, H., Corradini, S., Belka, C., Bogdanov, D., Boldrini, L., Bono, F., Goelz, C., Landry, G., Panza, G., Parodi, K., Talviste, R., Tran, H. E., Gambacorta, M. A., & Marschner, S. (2024). Privacy-friendly evaluation of patient data with secure multiparty computation in a European pilot study. *Npj Digital Medicine, 7*(1), 280. https://doi.org/10.1038/s41746-024-01293-4

Bladon, S., Eisner, E., Bucci, S., Oluwatayo, A., Martin, G. P., Sperrin, M., Ainsworth, J., & Faulkner, S. (2025). A systematic review of passive data for remote monitoring in psychosis and schizophrenia. *NPJ Digital Medicine, 8*, 62. https://doi.org/10.1038/s41746-025-01451-2

Bouguettaya, A., Stuart, E. M., & Aboujaoude, E. (2025). Racial bias in AI-mediated psychiatric diagnosis and treatment: A qualitative comparison of four large language models. *Npj Digital Medicine, 8*(1), 332. https://doi.org/10.1038/s41746-025-01746-4

Center for Devices and Radiological Health. (2025). *Artificial Intelligence and Machine Learning in Software as a Medical Device.* FDA. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Chandler, C., Foltz, P. W., & Elvevåg, B. (2022). Improving the Applicability of AI for Psychiatric Applications through Human-in-the-loop Methodologies. *Schizophrenia Bulletin, 48*(5), 949–957. https://doi.org/10.1093/schbul/sbac038

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithm fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering, 7*(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics, 6*. https://doi.org/10.3389/fhumd.2024.1421273

Choi, A., Ooi, A., & Lottridge, D. (2024). Digital Phenotyping for Stress, Anxiety, and Mild Depression: Systematic Literature Review. *JMIR mHealth and uHealth, 12*(1), e40689. https://doi.org/10.2196/40689

Clark, J. (2025, July 28). *AI Nutrition Labels—A Food-Inspired Approach To Trust—I.* IMD. https://www.imd.org/ibyimd/artificial-intelligence/ai-nutrition-labels-a-food-inspired-

approach-to-trust/

Clark, L. (2015, May 13). *Fitbit data could help schizophrenia sufferers avoid relapse*. Wired. https://www.wired.com/story/schizophrenia-relapse-alert-system-fitbit/

Colyer, A. (2020, February 18). The Way We Think About Data—ACM Queue. https://queue.acm.org/detail.cfm?id=3384393

Commissioner, O. of the. (2025, June 1). *FDA Issues Comprehensive Draft Guidance for Developers of Artificial Intelligence-Enabled Medical Devices*. FDA; FDA. https://www.fda.gov/news-events/press-announcements/fda-issues-comprehensive-draft-guidance-developers-artificial-intelligence-enabled-medical-devices

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health, 3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Definitive Healthcare. (2020, May 26). *How Does EHR Adoption Impact Data Sharing*? https://www.definitivehc.com/blog/hospital-ehr-adoption

Destiny, A. (2025). Leveraging Explainable AI and Multimodal Data for Stress Level Prediction in Mental Health Diagnostics. *International Journal of Research and Innovation in Applied Science, IX(XII)*, 416–425. https://doi.org/10.51584/IJRIAS.2024.912037

Ducharme, J. (2019, November 20). *Artificial Intelligence Could Be a Solution to America's Mental Health Crisis*. TIME. https://time.com/5727535/artificial-intelligence-psychiatry/

Garbazza, C., Mangili, F., D'Onofrio, T. A., Malpetti, D., Riccardi, S., Cicolin, A., D'Agostino, A., Cirignotta, F., Manconi, M., & "Life-ON" study group. (2024). A machine learning model to predict the risk of perinatal depression: Psychosocial and sleep-related factors in the Life-ON study cohort. *Psychiatry Research, 337*, 115957. https://doi.org/10.1016/j.psychres.2024.115957

Gerke, S. (2023). "*Nutrition Facts Labels*" for Artificial Intelligence/Machine Learning-Based Medical Devices—The Urgent Need for Labeling Standards.

Gilbert, S., Mathias, R., Schönfelder, A., Wekenborg, M., Steinigen-Fuchs, J., Dillenseger, A., & Ziemssen, T. (n.d.). A roadmap for safe, regulation-compliant Living Labs for AI and digital health development. *Science Advances, 11*(20), eadv7719. https://doi.org/10.1126/sciadv.adv7719

Golden, G., Popescu, C., Israel, S., Perlman, K., Armstrong, C., Fratila, R., Tanguay-Sela, M., & Benrimoh, D. (2023). *Applying Artificial Intelligence to Clinical Decision Support in Mental Health: What Have We Learned?* (No. arXiv:2303.03511). arXiv. https://doi.org/10.48550/arXiv.2303.03511

Golden, G., Popescu, C., Israel, S., Perlman, K., Armstrong, C., Fratila, R., Tanguay-Sela, M., & Benrimoh, D. (2024). Applying artificial intelligence to clinical decision support in mental health: What have we learned? *Health Policy and Technology, 13*(2), 100844. https://doi.org/10.1016/j.hlpt.2024.100844

Gumley, A. I., Bradstreet, S., Ainsworth, J., Allan, S., Alvarez-Jimenez, M., Birchwood, M., Briggs, A., Bucci, S., Cotton, S., Engel, L., French, P., Lederman, R., Lewis, S., Machin, M., MacLennan, G., McLeod, H., McMeekin, N., Mihalopoulos, C., Morton, E., … Gleeson, J. (2022). Digital smartphone intervention to recognise and manage early warning signs in schizophrenia to prevent relapse: The EMPOWER feasibility cluster RCT. *Health Technology Assessment (Winchester, England), 26*(27), 1–174. https://doi.org/10.3310/HLZE0479

Higgins, O., & Wilson, R. L. (2025). Integrating Artificial Intelligence (AI) With Workforce Solutions for Sustainable Care: A Follow Up to Artificial Intelligence and Machine Learning (ML) Based Decision Support Systems in Mental Health. *International Journal of Mental Health Nursing, 34*(2), e70019. https://doi.org/10.1111/inm.70019

Huang, X., Zhang, L., Zhang, C., Li, J., & Li, C. (2025). Postpartum Depression Risk Prediction Using Explainable Machine Learning Algorithms. *Frontiers in Medicine, 12*. https://doi.org/10.3389/fmed.2025.1565374

Hudon, A. (2025). A hybrid fuzzy logic–Random Forest model to predict psychiatric treatment order outcomes: An interpretable tool for legal decision support. *Frontiers in Artificial Intelligence, 8*. https://doi.org/10.3389/frai.2025.1606250

Itani, S., & Rossignol, M. (2020). At the Crossroads Between Psychiatry and Machine Learning: Insights Into Paradigms and Challenges for Clinical Applicability. *Frontiers in Psychiatry, 11*, 552262. https://doi.org/10.3389/fpsyt.2020.552262

Kelland, K. (2018, October 10). *Mental health crisis could cost the world $16 trillion by 2030*. Reuters. https://www.reuters.com/article/world/mental-health-crisis-could-cost-the-world-16-trillion-by-2030-idUSKCN1MJ2SG/

Kerz, E., Zanwar, S., Qiao, Y., & Wiechmann, D. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry, 14*. https://doi.org/10.3389/fpsyt.2023.1219479

Khalil, S. S., Tawfik, N. S., & Spruit, M. (2024). Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. *Patterns, 5*(7), 100990. https://doi.org/10.1016/j.patter.2024.100990

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 Healthcare: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. Biological Psychiatry. *Cognitive Neuroscience and Neuroimaging, 6*(9), 856–864. https://doi.org/10.1016/j.bpsc.2021.02.001

Mendes, J. P. M., Moura, I. R., Van de Ven, P., Viana, D., Silva, F. J. S., Coutinho, L. R., Teixeira, S., Rodrigues, J. J. P. C., & Teles, A. S. (2022). Sensing Apps and Public Data Sets for Digital Phenotyping of Mental Health: Systematic Review. *Journal of Medical Internet Research, 24*(2), e28735. https://doi.org/10.2196/28735

Mitrea, T., & Borda, M. (2020). Mobile Security Threats: A Survey on Protection and Mitigation Strategies. *International Conference Knowledge-Based Organization, 26*(3), 131–135. https://doi.org/10.2478/kbo-2020-0127

Ni, Y., & Jia, F. (2025). A Scoping Review of AI-Driven Digital Interventions in Mental Health Care: Mapping Applications Across Screening, Support, Monitoring, Prevention, and Clinical Education. *Healthcare, 13*(10), 1205. https://doi.org/10.3390/healthcare13101205

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), 100347. https://doi.org/10.1016/j.patter.2021.100347

Oudin, A., Maatoug, R., Bourla, A., Ferreri, F., Bonnot, O., Millet, B., Schoeller, F., Mouchabac, S., & Adrien, V. (2023). Digital Phenotyping: Data-Driven Psychiatry to Redefine Mental Health. *Journal of Medical Internet Research, 25*, e44502. https://doi.org/10.2196/44502

Pacheco, J., Garvey, M. A., Sarampote, C. S., Cohen, E. D., Murphy, E. R., & Friedman-Hill, S. R. (2022). The Contributions of the RDoC Research Framework on Understanding the Neurodevelopmental Origins, Progression and Treatment of Mental Illnesses. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 63*(4), 360–376. https://doi.org/10.1111/jcpp.13543

Pavez, J., & Allende, H. (2024). A Hybrid System Based on Bayesian Networks and Deep Learning for Explainable Mental Health Diagnosis. *Applied Sciences, 14*(18), Article 18. https://doi.org/10.3390/app14188283

Price, G. D., Heinz, M. V., Zhao, D., Nemesure, M., Ruan, F., & Jacobson, N. C. (2022). An unsupervised machine learning approach using passive movement data to understand depression and schizophrenia. *Journal of Affective Disorders, 316*, 132–139. https://doi.org/10.1016/j.jad.2022.08.013

Qiu, Y., Yao, H., Ren, P., Tian, X., & You, M. (2025). Regulatory sandbox expansion: Exploring the leap from fintech to medical artificial intelligence. *Intelligent Oncology, 1*(2), 120–127. https://doi.org/10.1016/j.intonc.2025.03.001

Rajashekar, N. (n.d.). *Generative Artificial Intelligence In Clinical Decision Support—Quantitative And Qualitative Analyses.*

Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science, 19*(1), 27. https://doi.org/10.1186/s13012-024-01357-9

Ren, W., Liu, Z., Wu, Y., Zhang, Z., Hong, S., & Liu, H. (n.d.). Moving Beyond Medical Statistics: A Systematic Review on Missing Data Handling in Electronic Health Records. *Health Data Science, 4*, 0176. https://doi.org/10.34133/hds.0176

Rodriguez, T., MA, & LPC. (2025, June 13). *Addressing Racial Bias in Pulse Oximeter Accuracy*. The Cardiology Advisor. https://www.thecardiologyadvisor.com/features/pulse-oximeter-accuracy-racial-bias/

Sadeh-Sharvit, S., & Hollon, S. D. (2025). *AI Integration in Behavioral Healthcare: A Practical Framework for Clinicians.* Journal of Technology in Behavioral Science. https://doi.org/10.1007/s41347-025-00532-z

Shaik, T., Tao, X., Xie, H., Li, L., Higgins, N., & Velásquez, J. D. (2025). Towards Transparent Deep Learning in Medicine: Feature Contribution and Attention Mechanism-Based Explainability. *Human-Centric Intelligent Systems, 5*(2), 209–229. https://doi.org/10.1007/s44230-025-00104-7

Son, C., Hegde, S., Markert, C., Zahed, K., & Sasangohar, F. (2023). Use of a Mobile Biofeedback App to Provide Health Coaching for Stress Self-management: Pilot Quasi-Experiment. *JMIR Formative Research, 7*(1), e41018. https://doi.org/10.2196/41018

Spoelma, M. J., Serafimovska, A., & Parker, G. (2023). *Differentiating melancholic and non-melancholic depression via biological markers: A review.* The World Journal of Biological Psychiatry. https://www.tandfonline.com/doi/abs/10.1080/15622975.2023.2219725

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry, 10*(1), 116. https://doi.org/10.1038/s41398-020-0780-3

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *Npj Digital Medicine, 3*(1), 17. https://doi.org/10.1038/s41746-020-0221-y

Ul Hussna, A., Immami Trisha, I., Jahan Ritun, I., & Rabiul Alam, Md. G. (2021). COVID-19 impact on students' Mental Health: Explainable AI and Classifiers. 2021 International Conference on Decision Aid Sciences and Application (DASA), 847–851. https://doi.org/10.1109/DASA53625.2021.9682371

Vafaei Sadr, A., Li, J., Hwang, W., Yeasin, M., Wang, M., Lehmann, H., Zand, R., & Abedi, V. (2025). Flexible imputation toolkit for electronic health records. *Scientific Reports, 15*(1), 17176. https://doi.org/10.1038/s41598-025-02276-5

Van Der Donckt, J., Vandenbussche, N., Van Der Donckt, J., Chen, S., Stojchevska, M., De Brouwer, M., Steenwinckel, B., Paemeleire, K., Ongenae, F., & Van Hoecke, S. (2024). Mitigating data quality challenges in ambulatory wrist-worn wearable monitoring through analytical and practical

approaches. *Scientific Reports, 14*(1), 17545. https://doi.org/10.1038/s41598-024-67767-3

Wang, G., Bennamoun, H., Kwok, W. H., Quimbayo, J. P. O., Kelly, B., Ratajczak, T., Marriott, R., Walker, R., & Kotz, J. (2025). Investigating Protective and Risk Factors and Predictive Insights for Aboriginal Perinatal Mental Health: Explainable Artificial Intelligence Approach. *Journal of Medical Internet Research, 27*(1), e68030. https://doi.org/10.2196/68030

Wen, L.-Y., Zhang, L., Zhu, L.-J., Song, J.-G., Wang, A.-S., Feng, Y., Tao, Y.-J., Zhu, Y., Jin, Y.-L., & Chang, W.-W. (2025). Latent class analysis on mental health and associated factors in medical and non-medical college students. *Journal of Affective Disorders, 388*, 119593. https://doi.org/10.1016/j.jad.2025.119593

WHO. (2025). Mental health. World Health Organization. https://www.who.int/health-topics/mental-health

Wim, V. (2025, June 27). *The EU AI Act and Medical Devices: Navigating High-Risk Compliance, Wim Vandenberghe*. Reed Smith LLP. https://viewpoints.reedsmith.com/post/102kq35/the-eu-ai-act-and-medical-devices-navigating-high-risk-compliance

Zhang, Y., Wang, J., Zong, H., Singla, R. K., Ullah, A., Liu, X., Wu, R., Ren, S., & Shen, B. (2025). The comprehensive clinical benefits of digital phenotyping: From broad adoption to full impact. *Npj Digital Medicine, 8*(1), 196. https://doi.org/10.1038/s41746-025-01602-5

Zhu, H., Bai, J., Li, N., Li, X., Liu, D., Buckeridge, D. L., & Li, Y. (2025). FedWeight: Mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. *Npj Digital Medicine, 8*(1), 286. https://doi.org/10.1038/s41746-025-01661-8

Zulqarnain, M., Shah, H., Ghazali, R., Alqahtani, O., Sheikh, R., & Asadullah, M. (2023). Attention Aware Deep Learning Approaches for an Efficient Stress Classification Model. *Brain Sciences, 13*(7), 994. https://doi.org/10.3390/brainsci13070994