International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

AI-Driven Mental Health Intelligence and Early Intervention Framework: An Enterprise Platform Architecture Perspective

Muneer Basha Shaik[1] and Anitha Velde[2]

Abstract: With 1.17 billion people worldwide with mental health issues, there is a meaningful gap between clinical demand and the availability of production-grade smart monitoring systems. We present an enterprise platform architecture for AI-based mental health intelligence and early intervention systems, from the standpoint of large-scale distributed systems engineering, rather than a new class of clinical algorithms. The architecture is based on event-driven multimodal data pipelines, LLM reasoning layers with retrieval-augmented generation (RAG), federated learning, and a Digital Mental Health Twin construct that creates a continuously updated behavioral and psychological profile per individual. Explainability components help clinicians to see the reasoning behind outputs, while retaining human diagnostic authority. The architecture allows for early recognition of risk. Edge-cloud hybrid deployment patterns address the latency and data-sensitivity constraints of mental health applications. The holistic architecture is qualitatively evaluated against common non-functional requirements missing in existing health AI proposals: scalability, compatibility with privacy regulations, interpretability, and clinical trust. As an enterprise integration problem, the architecture acts as a deployable framework that healthcare organizations, workplace wellness programs, schools, and veteran services may implement to scale AI-informed preventive mental healthcare in a manner compatible with the culture of practice.

Keywords: *Multimodal AI Integration, Mental Health Monitoring, Enterprise Platform Architecture, Event-Driven Processing, Federated Learning, Explainable AI*

1. Introduction

One of the main public health challenges of the decade is mental health disorders. The 2023 Global Burden of Disease Study estimates that 1.17 billion people have mental illness, nearly double the 599 million cases when calculated in 1990 [1]. Depression alone affects 280 million and globally anxiety disorders affect 301 million or 4.05% of the population and are the most common cause of disability worldwide [2, 3]. The economic cost of depression and anxiety disorders is likewise high. They cost the global economy approximately US$1 trillion per year in lost productivity. Without a concerted effort to increase treatment coverage, an estimated 12 billion working days per year will be lost worldwide through 2030 due to depression and anxiety disorders [4].

Despite this scale, delivery of mental healthcare continues to be structurally reactive, with current clinical strategies for engagement and assessment based on periodic assessments, patient-initiated contact and episodic assessment, all of which are poor at recognizing a gradual process of behavioral decline. The treatment gap is particularly large in low and middle income countries, in which 76 per cent to 85 per cent of people with mental disorders do not receive any treatment, and in high income countries where, even in well-resourced health systems, months or years may intervene before clinical contact occurs.

The engineering question is not how to build better algorithms, but how to build production-grade platforms that can ingest, correlate and reason over heterogeneous behavioral signals at population scale, while also fulfilling the privacy guarantees, regulatory compliance posture and clinical explainability requirements of real world healthcare deployments. While promising machine learning models have been built in the literature, which detect behavioral risk factors based on speech, sensor and interaction data [6, 7], what is not systematically addressed is the enterprise integration

[1,2]*Independent Researcher, USA*

architecture required to deploy such models in real healthcare and organizational settings.

We develop and present a platform architecture for AI-enabled mental health intelligence and early intervention. The architectural contribution of this paper is to compose event-driven streaming pipelines, LLM-based reasoning with retrieval-augmented generation, federated learning, a Digital Mental Health Twin construct, and an explainability governance layer into a composable, deployable and scalable mental health intelligence platform. This analysis, following the enterprise platform architecture practice, considers the design decisions around the non-functional requirements (latency, throughput, privacy, fairness, interpretability) that separate production deployment from that of a research prototype.

The rest of this paper is organized as follows: In Section 2, we discuss the architecture and system overview. Section 3 describes how multimodal data are ingested and processed in a stream; section 4 describes LLM reasoning technologies with RAG; section 5 describes the federated learning and privacy-preserving intelligence layer; and section 6 describes the Digital Mental Health Twin construct. Section 7 discusses explainability, clinical trust and governance. Section 8 covers deployment architecture and infrastructure optimization, and Section 9 covers the larger implications and contexts of deployment. Section 10 concludes.

2. Architectural Paradigm and System Overview

The high-level architecture contains four architectural layers: a multimodal data ingestion and stream processing layer, an intelligence layer implementing LLM reasoning and predictive analytics, a longitudinal modeling layer implementing the Digital Mental Health Twin, and a clinical delivery and governance layer implementing explainable insights to clinical and administrative users. These layers are interconnected via an event-driven messaging fabric, and each layer is implemented as a separately deployable and horizontally scalable service.

The microservices-first approach is motivated by the diversity of the data produced for mental health monitoring. For example, wearable sensors can produce data at sub-second rates while notes recorded by clinicians may be produced far less often - in some cases over multiple weeks. A monolithic architecture would find it difficult if not impossible to optimize the system to handle the ingestion throughput of streaming data which must also be semantically understood. By separating each of the components into separate services, each can be independently scaled depending on the current need, and improved independently as models change.

The architecture is cloud-native and uses container orchestration systems for service lifecycles, autoscaling and fault isolation. The services are stateless and all persistent state is externalized to the storage systems appropriate for purpose. Time-series databases are used for the physiological signals data, vector databases for semantic embeddings and structured relational databases for longitudinal clinical records. This separation of compute and state is also the key to the flexible scaling that population-level deployment requires.

This architecture is meant to adopt a different approach than the predominant architecture used by EHR systems, where behavioral signals are recorded at specific points in time (e.g. six-month discharge summary, repeated PHQ-9 score, clinical note at intake, etc.), but are not continuously monitored between encounters. In contrast, the architecture treats behavioral signals as streams, enabling detection of changepoints in trend over time. Thus, the impact on the clinical workflow will not be to replace the clinician, but to provide a longitudinal behavioral context that no human practitioner could maintain across a population of patients.

Security and interoperability are accomplished via standards-compliant integration adapters connecting electronic health record systems using HL7 FHIR APIs, telehealth applications using secure streaming endpoints and device networks connected using an IoT gateway architecture. The architecture utilizes zero-trust principles to implement mutual authentication, encryption of data at rest and in transit, and role-based access control of data flow between each service.

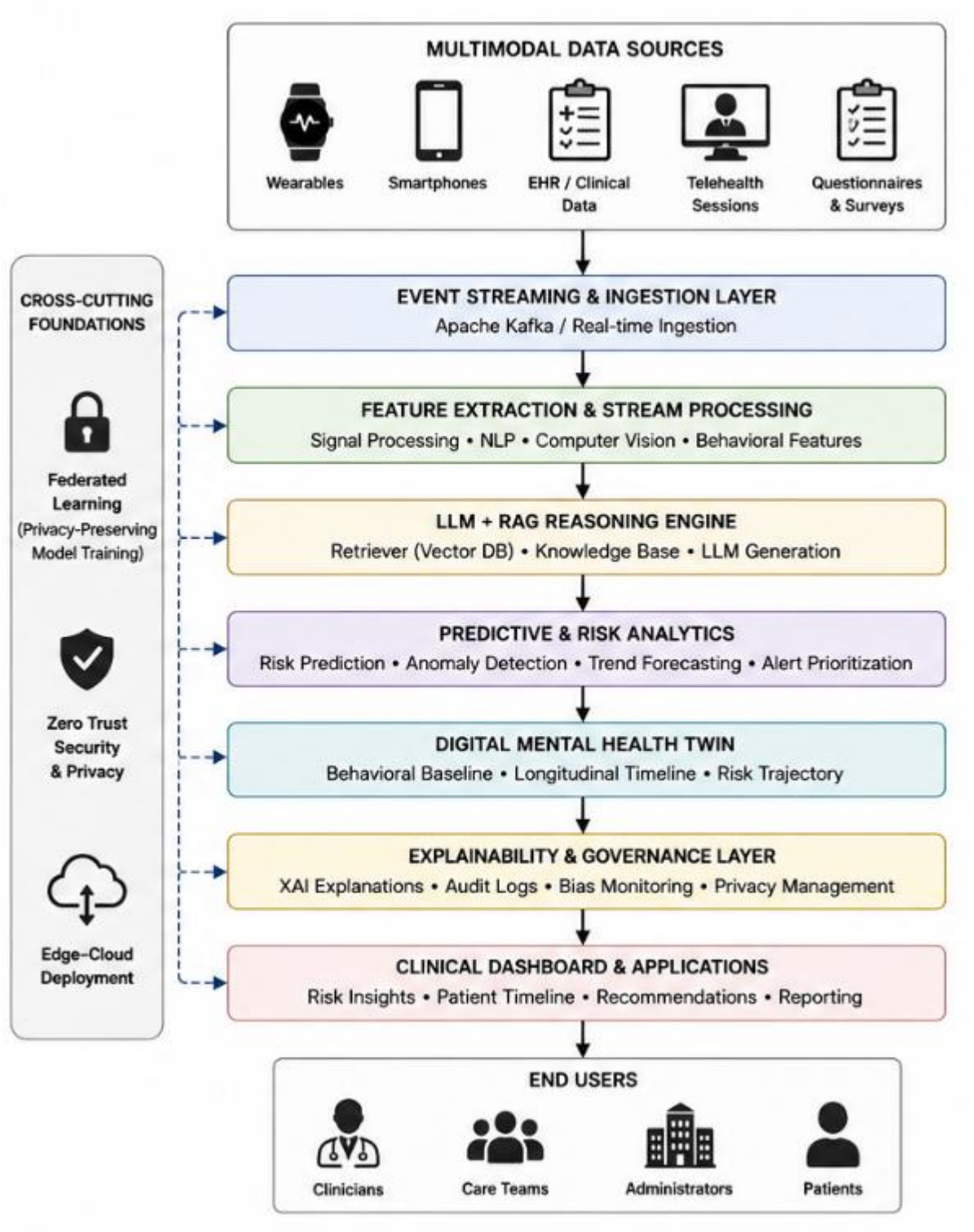

Figure 1. Overall Enterprise Architecture for AI-Driven Mental Health Intelligence and Early Intervention Framework

3. Multimodal Data Ingestion and Stream Processing

Such a system's intelligence depends on the amount, diversity, and recency of the data it is trained with. Various aspects of human behavior that are affected by mental health disorders can be tracked by speech and prosody features, activity patterns from wearables' accelerometers, sleep structure from heart rate variability and motion sensors, social activity from communication frequency and linguistic features from text. [7] The discriminative power of the single modality itself may not be

sufficient to classify a given change; instead, the correlation of changes across modalities is used.

The ingestion tier is built around an event streaming architecture that queues streams of data from wearables, mobile health apps, clinical information systems, and telehealth applications. Each data source writes to dedicated topic partitions within the event streaming architecture to allow ingestion of data streams at varying rates while decoupling data producers from consumers. This architecture prevents a high-velocity source (e.g. wrist-mounted gyroscope sampling in the hundreds of hertz) from causing backpressure that would delay the processing of lower-frequency clinical data.

The first processing stage (immediately downstream of the ingestion), called feature extraction, uses modality-specific signal processing to derive acoustic features from the voice, segment activity from accelerometer streams, estimate sleep stages from physiological sensor fusion, and encode linguistic features from the textual input. The architecture must be strong to noise. Since naturalistic data is often noisier, ingestion pipelines generally need adaptive filtering and quality-scoring mechanisms that flag and attenuate mistakes in this data rather than propogating artifacts into the intelligence layer.

Dimensionality reduction is performed as required during feature extraction to include only clinically relevant signals, while minimizing computational costs for downstream models. All feature data are appended with provenance metadata, including source device, collection timestamp, quality score, and individual context identifier, before being written to the intelligence tier. This metadata is required for the longitudinal modeling component, as to check for an anomaly, the current feature vector needs to be compared against a baseline, which can vary greatly between individuals.

The technical feasibility of multimodal sensing is verified in literature. According to Torous et al. passive sensing data (e.g., accelerometry, GPS displacement, tone of voice, and device usage patterns) from smartphones and wearables can provide behaviorally meaningful mental health data with acceptable patient burden [6]. Vázquez-Romero and Gallardo-Antolín have demonstrated that ensemble CNNs applied to speech acoustic features can detect and classify depressive speech patterns, a potential benefit for the ingestion pipeline's voice-based feature extraction capabilities [8]. Thus, this architecture is designed to accept ensembles of this type as a plug-and-play feature extraction service allowing models to be deployed independently of the pipeline infrastructure.

Table 1.Multimodal Data Sources and AI Processing Components

Data Source	Extracted Features	AI Processing Component	Clinical Purpose
Wearable Sensors	Heart rate, activity, sleep	Feature Extraction Engine	Behavioral monitoring
Smartphone Usage	Screen activity, mobility, communication	Event Processing Pipeline	Lifestyle pattern analysis
Speech Recordings	Prosody, pitch, pauses	CNN/Acoustic Analysis	Depression detection
Clinical Notes	Diagnoses, observations	LLM + RAG	Contextual reasoning
Patient Questionnaires	PHQ-9, GAD-7 scores	Predictive Analytics	Risk assessment
Telehealth Sessions	Conversation transcripts	NLP + LLM	Behavioral trend analysis

4. LLM-Powered Reasoning and RAG Integration

Language models would be used for contextual reasoning over heterogeneous inputs, rather than for behavioral signal classification as smaller task-specific models are better than a generalist language model at this task. Clinical notes, patient-reported outcomes and therapeutic process observations, as well as transcripts of therapy conversations are difficult to vectorize, but a language model could produce structured risk assessments, identify

discrepancies, and surface contextual information relevant to decision-making by clinicians.

Naively inserting a large language model into a clinical reasoning pipeline is problematic because large language models trained on general-purpose datasets could issue plausible but clinically incorrect statements. This is more problematic in mental health settings, because misplaced optimism could delay appropriate care. The LLM utilizes a retrieval-augmented generation pipeline, where the LLM's reasoning is based on information dynamically retrieved from multiple knowledge sources such as clinical practice guidelines, evidence summaries, anonymized population-level pattern libraries, and the person's own longitudinal behavioral records.

The workflow of the RAG architecture is the following. A reasoning task is triggered when an alert fires from the predictive risk model, a clinical question is posed, or a longitudinal review has to be conducted, and a question-specific embedding of the current context is obtained and fed into the vector database. Retrieved data is injected into the context window of the model, along with the structured behavioral summary created by the upstream analytics. The model then produces a reasoning output restricted to the evidence provided, citing passages from the retrieved documents. Such a grounding mechanism will greatly reduce the rate of unsupported assertion and provide an auditable basis for every model output.

This tiered model strategy has architectural implications in deploying lightweight distilled models for high-frequency, low-complexity reasoning tasks like initial risk scoring, summarizing sentiment trends, and producing alerts. The larger foundation models are only used for lower frequency, deeper reasoning tasks (holistic longitudinal profiling, multi-factorial risk rationale, clinician natural language question-answering). This prevents too many calls to resource-intensive models and also allows the architecture and use of foundation models to be reflective of the depth or level of reasoning being performed.

A key guard rail for the integration of LLMs into clinical workflows is that they should not provide independent clinical recommendations. This is accomplished via a human-in-the-loop design pattern in which model predictions are passed through a clinical delivery layer that produces structured decision-support recommendations with uncertainty estimates and retrieval model citations with human clinician review and validation. The model cannot initiate clinical action by itself and merely informs human clinical decision making. This is enforced both ethically and architecturally via service interfaces.

5. Federated Learning and Privacy-Preserving Intelligence

Centralized model training is the customary method of training a machine learning model. However, this entails shipping all of the training data to a central server which is contrary to even the most minimal privacy requirements for psychiatric information. Mental health data, being sensitive with strict data-handling regulations (such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, and equivalent laws in other jurisdictions), cannot ethically or legally be aggregated in one central repository.

Federated learning resolves this paradox by inverting the data movement model (model to data instead of data to model). Each local node (the healthcare institution, an enterprise wellness installation, or a device client) learns model updates locally, using data it stores locally, and then communicates gradients of model parameters to a centralized aggregation server. The aggregation server employs algorithms like FedAvg to update the global model post-aggregation, before disseminating it back to each local node. At no point is the raw individual data sent out of its custodian's organizational boundary [9].

An additional privacy solution can be applied in form of differential privacy by adding noise to the gradient sent to the model. This allows for a formal privacy guarantee regarding the leakage of the local user's data depending on the global model. In this case, differential privacy is maintained in the local round (i.e., by gradient clipping and noise addition). In practice, different privacy budget parameters can be set depending on the deployment use case to offset the accuracy/privacy trade-off.

Several challenges need to be overcome in federated learning architectures, one of which is communication efficiency: sharing gradients is expensive, especially in the case of large models. To this end, it outlines gradient compression, sparse update transmission, and asynchronous aggregation protocols tolerant of nonconstant node participation rates, as would be expected in practice (e.g., due to

off-line maintenance for healthcare institutions or varying numbers of patients enrolled at different points in time).

Models need to be controlled in federated settings to prevent degraded performance from heterogeneous local datasets, since federated nodes can have different population distributions that contribute model updates that reduce the global model's performance on other populations. Additionally, contribution quality scoring and anomaly detection in the aggregation layer allow the system to down-weight or exclude contributors with anomalous gradient contributions. This is of particular relevance in mental health applications, where the behavior signatures of mental health conditions may vary substantially between the cultures, demographics and socioeconomic groups of the contributors.

6. Digital Mental Health Twin and Longitudinal Modeling

The Digital Mental Health Twin is a longitudinal modeling construct that forms the conceptual core of the proposed framework. Existing methods of capturing the mental health profile of a patient rely on discrete clinical snapshots. The Digital Mental Health Twin is an evolving approximate representation of a person's mental and behavioral state. Factors which are not captured through a single metric must be compared to an individual-specific baseline representation and depend upon how the individual's baseline behavior changes over time [10].

Multiple layers of longitudinal information make up each Digital Mental Health Twin. A baseline behavioral profile is the first layer and is formed during a first observation period. This profile represents their baseline across all modalities being monitored, including sleep schedule and architecture, activity, communication, voice prosody distributions, and linguistic sentiment distributions. These new baselines are compared to previously established baselines to assess changes in the individual's status. The deviation-detection component is a symmetric mixture of statistical process control and ML-based anomaly detection methods, tuned around a balance between maximizing deviation detection and minimizing false positive rates, an important design constraint because too many false positives could reduce clinician trust of the system.

More generally, the Digital Mental Health Twin can help evaluate the effects of interventions (medications, psychotherapies, behavioral activation and organizational support). When an intervention is started, the digital twin's longitudinal record of baseline behavior is available as a point against which behavioral change can be quantitatively assessed following an intervention. The platform addresses a long-standing issue in mental healthcare: namely, that no objective measures have been available between clinical visits to gauge a patient's response to treatment. The dashboard/interface allows clinicians to see how a person's behavioral indicators trend toward or away from a normative baseline after an intervention.

The Digital Mental Health Twin can be deployed at industrial scale with suitable computational architecture. In an enterprise wellness application, there are tens of thousands of employees, and in a healthcare system, hundreds of thousands of patients: Each individual requires a continuously updated probabilistic model, which requires a scalable architecture. A scalable architecture can be achieved by partitioning the behavioral data storage between recent behavioral windows in a high-performance time-series database and historical data in a low-cost object store, decoupling real-time monitoring from periodic recomputation via asynchronous model update pipelines, and implementing fine-grained twin activation that is triggered via lower-frequency updates in low-risk individuals and higher-frequency updates in cases of detected anomalies.

Additionally, the twin construct allows for adaptive intervention planning. The API exposes trend data over time so that care coordinators and clinicians can assess risk by predicting when an individual's behavior may reach a risk threshold by measuring the direction and velocity of an observed trajectory. The goal of this capability is to shift the clinical workflow from case management to risk navigation.

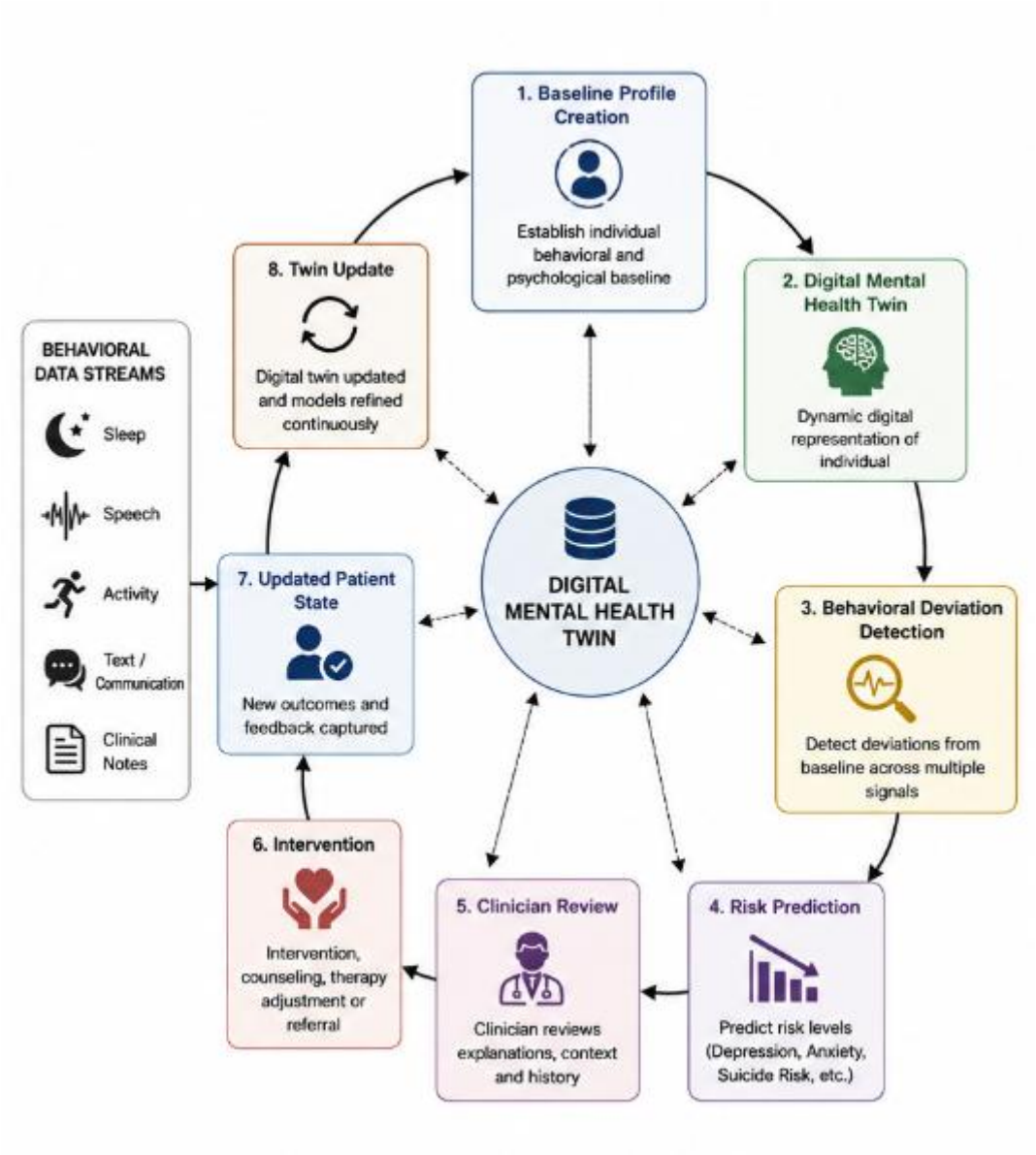

Figure 2. Digital Mental Health Twin Lifecycle and Continuous Learning Workflow

7. Explainability, Clinical Trust, and Governance

Understanding a clinical AI system's reasoning is a requirement, not just a desirable property. Clinicians often cannot use clinical judgment to act on an AI's recommendation if they cannot interrogate the AI's reasoning. Regulators cannot be assured a system is not making prohibited recommendations if its reasoning cannot be understood, and patients cannot consent to being guided by medical advice they cannot comprehend. It therefore treats explainability as a first-class architectural concern, instead of a post-hoc product feature.

Next, the explainability layer provides two different kinds of explanations for the model's ranking: at the feature attribution level, it produces a set of factor importance scores to communicate which behavioral signals had the highest impact on a risk prediction. Finally, these attributions are calculated using SHAP decomposition methods [11], which provide model-agnostic, locally accurate estimates of the importance of different features. When using the system's dashboard, the clinician sees a ranked list of elements contributing to a given score, including, for an individual risk score, a lower estimate of sleep regularity, a lower estimate of conversational frequency or an increased estimate of linguistic

markers in the conversation that indicate cognitive load.

The reasoning transparency level is achieved by citing an LLM's reasoning in terms of the retrieved documents and behavioral traces. With such a citation strategy, clinicians can easily retrace any model assertion back to its supporting evidence, validate the evidence, and ascertain whether the model relied on unrepresentative or irrelevant context to arrive at a belief or behavior. At the architectural level, low-confidence model outputs that cannot be ground-truthed from retrieved evidence are sent to a manual reviewer process before they appear in clinical interfaces.

The governance layer incorporates bias detection and fairness monitoring. This type of bias has been shown to be present in predictive algorithms in health management when analyzed by population subgroups. Obermeyer et al. showed that for a widely adopted health risk algorithm, compared to an equally sick low-risk population, these algorithms under-identify the high need patients belonging to that population subgroup by 49% [12]. The architecture addresses this using continuous fairness auditing. The performance across the subgroups is computed separately, and the governance pipeline alerts of meaningful subgroup discrepancies so that a human operator can intervene, and it rejects new model updates that create or worsen subgroup fairness discrepancies.

At both the interface and governance levels, the AI system is treated as a decision support system. For the former, the system shows to the user risk indicators and contributing factors rather than diagnosis and treatment recommendations. For the latter, all model outcomes, retrieved evidence items, and human responses are recorded in audit trails. These audit trails serve to document regulatory compliance and provide the basis for quality assurance review.

8. Deployment Architecture and Infrastructure Optimization

A key operational tension is that real-time or time-critical tasks such as anomaly detection and alert generation should have sub-second latency, while other batch tasks with more computational intensity such as federated model aggregation, longitudinal twin recomputation and analytics can be performed on a less than fully-loaded system. Infrastructure should support workload-aware resource orchestration that dynamically schedules resources based on clinical priority when required, optimizing for the appropriate computational resource.

An edge-cloud hybrid architecture is often preferred, where data preprocessing, lightweight inference for common patterns in risk scoring, and sensitive operations that should not transmit raw data to the cloud, are done on the edge (locally on-device, on-premise, or at the organizational gateway). As a result only aggregated feature vectors, encrypted intermediate representations, and federated gradient updates are transmitted to the cloud infrastructure, which reduces the data transmission latency, and the surface area over which sensitive data may leak or be exposed.

Container orchestration manages the cloud infrastructure. Autoscaling policies create and destroy replicas of services based on queue depth and processing latency. Service mesh configurations will often include features such as circuit breakers and retries to prevent the service from going down should a service or component fail. This is important in a clinical system, where failures may put patients at risk. Canary deployment patterns can be used to validate model updates or infrastructure changes on a small fraction of traffic prior to full deployment, thus minimizing risk of regressions in production.

The different read/write characteristics of mental health monitoring data lends itself to tiered data storage, where recent behavioral windows (i.e., sensor data, communication network data, and assessment scores) of seven to thirty days are stored in high performance time-series databases optimized for range queries on individual identifiers and time dimensions. Historical data is stored to object storage at a much lower cost per gigabyte. The data is available for retrospective analysis and model training but has higher latency that is best suited to non-real-time applications.

Energy and cost are managed by scheduling workloads to cloud provider, computing pricing and carbon intensity profiles. Batch workloads (federated aggregation cycles, longitudinal twin recomputation, population-level analytics) are scheduled based on cloud provider pricing and grid carbon intensity profiles, optimizing for cost and carbon emissions for the computation to be executed. Real-time monitoring and alerting pipelines are exceptions to this constraint, due to urgent clinical requirements. Runtime telemetry

allows service resource consumption to be right-sized through feedback into the capacity planning system as a service scales up or down with use.

Table 2.Deployment Architecture and Infrastructure Optimization

Architectural Component	Primary Function	Enterprise Benefit
Event Streaming Layer	Continuous multimodal ingestion	High throughput
Feature Extraction	Signal preprocessing	Reduced computational cost
LLM + RAG Engine	Evidence-grounded reasoning	Explainability
Federated Learning	Distributed model training	Privacy preservation
Digital Mental Health Twin	Longitudinal behavioral modeling	Personalized monitoring
Explainability Layer	SHAP-based interpretation	Clinical trust
Governance Layer	Bias detection and auditing	Regulatory compliance
Edge–Cloud Deployment	Hybrid processing	Low latency & scalability

9. Societal Implications and Deployment Contexts

Requirements for cloud-native deployment, API-based integration, mobile data collection, and privacy preserving intelligence give the proposed platform the potential for deployment outside of customary healthcare institutions. Each deployment scenario raises its own unique set of requirements, stakeholders, and ethical considerations, and this heterogeneity needs to be accounted for in the platform architecture.

In workplace wellness deployments, the framework eases the evaluation of population level mental health trends (aggregate burnout scores, distribution of trend scores for stress, movements in engagement patterns), without exposing individual level identifiable information. Based on the aggregated data, organizations can assess whether their workload, leave policies, and wellness resources are appropriate. This individual-level monitoring capability is still possible under the voluntary enrollment model if coupled with opt-in consent and data use disclosures. Rajkomar et al. noted that machine learning applied to health data can surface clinically meaningful patterns at population scale [13]. It is then proposed to integrate this with occupational mental health.

In the case of an educational setting and if the student is a minor, the consent management and data governance components of the architecture should be configurable in order to ease parental consent and compliance with obligations around data minimization and age-appropriate disclosures. Behavior monitoring in the educational context is generally less sensitive, and is focused on academic engagement, communication data, and reported wellbeing.

Longitudinal modeling has been used in veteran support services. Veterans that are seeking support services may experience complex patterns of accumulation and recovery from PTSD, depression, and adjustment disorders. Moreover, the DMHT allows tracking and comparison of the longitudinal trajectories of these symptoms over time, which is particularly relevant to longitudinal care provided to veterans. The DMHT's use alongside existing veteran health information systems may raise additional data governance considerations, such as the security classification of service history data co-located with mental health data.

In all deployment contexts, the social value of the framework depends on responsible deployment. AI ethics that consider automated systems (including health AI) document how without practices that consider the demographics of the affected population and the distribution of model performance across groups, social inequities can be exacerbated by AI design [14, 15]. In addition to bias monitoring and bias governance, the framework also requires (1) stakeholder engagement with affected communities, (2) public disclosure of the capabilities, limitations, and performance characteristics of a system, and (3) meaningful mechanisms for individuals to contest or seek review for decisions made based on the outputs of a model.

10. Conclusion

The magnitude of the mental ill-health global burden (1.17 billion people affected, annual productivity loss of a trillion dollars, and an estimated treatment gap of greater than 76% in low-resource settings) makes a compelling case for ensuring that technology-enabled early detection and intervention is possible. A more challenging but equally key factor not well articulated in the literature is what enterprise platform architecture is needed for AI-based mental health monitoring to be scalable and operationally viable. In this paper, we have proposed this architecture consisting of event-driven multimodal ingestion pipelines, language models for reasoning with retrieval-augmented generation, federated learning for privacy-preserving intelligence, the Digital Mental Health Twin longitudinal modeling construct, and a thorough explainability and governance layer.

Technical architectures are needed that meet functional constraints that distinguish production deployment from research prototype (sub-second delay for crisis detection, regulatory compliance on privacy, clinical interpretability to retain practitioner trust, and fairness monitoring to prevent differential treatment). Algorithms alone cannot meet these requirements. This requires purposefully architecting and engineering these systems-as-a-service, a standard form for mainstream enterprise platform engineering of financial, logistics, communications, and social systems, but which the health AI research literature has not yet revisited and systematically adapted for mental health applications.

Given that this is a framework paper, here we provide a qualitative evaluation against the requirements. Future work would include a controlled evaluation of the components against established benchmarks: the performance of the RAG-grounded reasoning layer against clinical assessment benchmarks, the convergence speed of the federated-learning training compared with a centralized-training baseline, and the performance of the Digital Mental Health Twin anomaly detection against longitudinal clinical outcome data. Experience shows us that there are deployment-specific constraints in live healthcare environments, which cannot be captured by architectural proposals.

Reconceptualizing active clinical mental health monitoring as an enterprise systems integration challenge makes the engineering work needed to close the divide between clinical AI research and large-scale implementation more visible. This gap is clearly not algorithmic, as effective models for behavioral risk detection already exist and continue to improve over time. Architectural, it is the gap that this framework is designed to fill.

Ethics Statement

This paper proposes a conceptual framework and does not report empirical research involving human participants. No primary data collection, patient interaction, or clinical experimentation was conducted in preparing this work. All referenced statistics are derived from previously published, peer-reviewed sources. Authors declare no conflicts of interest.

References

[1] GBD 2023 Mental Disorder Collaborators, "Updated trends in the global prevalence and burden of mental disorders, 1990–2023: a systematic analysis for the Global Burden of Disease Study 2023," The Lancet, vol. 407, no. 10543, pp. 2040–2064, May 2026. DOI: 10.1016/S0140-6736(26)00519-2.

[2] A. A. Hashim et al., "Epidemiology of anxiety disorders: global burden and sociodemographic associations," Middle East Curr. Psychiatry, vol. 30, no. 1, p. 38, 2023. DOI: 10.1186/s43045-023-00315-3.

[3] World Health Organization, "Mental disorders," WHO Fact Sheet, Sep. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[4] D. Chisholm et al., "Scaling-up treatment of depression and anxiety: a global return on investment analysis," Lancet Psychiatry, vol. 3, no. 5, pp. 415–424, 2016. DOI: 10.1016/S2215-0366(16)30024-4.

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

[6] J. Torous et al., "The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual

reality," World Psychiatry, vol. 20, no. 3, pp. 318–335, 2021. DOI: 10.1002/wps.20883.

[7] M. Sheikh, M. Qassem, and P. A. Kyriacou, "Wearable, environmental, and smartphone-based passive sensing for mental health monitoring," Front. Digit. Health, vol. 3, p. 662811, 2021. DOI: 10.3389/fdgth.2021.662811.

[8] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," Entropy, vol. 22, no. 6, p. 688, 2020. DOI: 10.3390/e22060688.

[9] P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," Found. Trends Mach. Learn., vol. 14, no. 1–2, pp. 1–210, 2021. DOI: 10.1561/2200000083.

[10] S. Elkefi and O. Asan, "Digital twins for managing health care systems: rapid literature review," J. Med. Internet Res., vol. 24, e37641, 2022. DOI: 10.2196/37641.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.

[12] Z. Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019. DOI: 10.1126/science.aax2342.

[13] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," npj Digital Med., vol. 1, no. 18, 2018. DOI: 10.1038/s41746-018-0029-1.

[14] A. Jobin et al., "The global landscape of AI ethics guidelines," Nat. Mach. Intell., vol. 1, pp. 389–399, 2019. DOI: 10.1038/s42256-019-0088-2.

[15] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," JAMA, vol. 319, no. 13, pp. 1317–1318, 2018. DOI: 10.1001/jama.2017.18391.

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

[17] T. Insel, "Digital phenotyping: technology for a new science of behavior," JAMA, vol. 318, no. 13, pp. 1215–1216, 2017. DOI: 10.1001/jama.2017.11295.

[18] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), 2017, pp. 1273–1282.

[19] M. T. Ribeiro et al., "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[20] National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF 1.0)," NIST, Gaithersburg, MD, 2023. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-1.