Federated AI for Mental Health: A Privacy-Preserving GAN-LLM Framework for Risk Prediction and Early Detection

NP Thothela[1*] and A Bagula[2]

[1] Central University of Technology, Bloemfontein, South Africa

2 University of the Western Cape, Cape Town, South Africa
portia.thothela@gmail.com

Abstract. Mental health disorders affect nearly 30% of South Africa's population, yet access to adequate care remains significantly limited due to socioeconomic barriers, stigma, and resource shortages. This study presents Federated GANS-LLM, a novel hybrid framework that combines federated learning (FL), generative adversarial networks (GANs), and large language models (LLMs) to improve mental health risk prediction, facilitate early detection, and enhance patient-centered care while maintaining strict data privacy compliance. The proposed framework is built on a multi-layered architecture comprising three key components. The Client-Side Federated Learning layer enables healthcare institutions to process patient data locally, transmitting only encrypted model updates to ensure privacy protection. The GAN-Based Data Augmentation layer includes a centralized GAN module enhanced with variational autoencoders (VAE) to generate high-quality synthetic data, addressing data sparsity and heterogeneity to improve model robustness. The ClinicalBERT-Driven Risk Assessment layer employs an advanced LLM that analyzes contextual indicators, sentiment, and linguistic markers to assess mental health risks with high interpretability. By assessing the technical feasibility, ethical considerations, and real-world impact of this approach, the research contributes to the advancement of scalable, privacy-preserving AI solutions for addressing global mental health challenges. This work also demonstrates the transformative potential of AI-driven, privacy-focused innovations in healthcare, positioning federated learning as a viable approach to bridging the mental health treatment gap worldwide.

Keywords: Federated Learning (FL), Generative Adversarial Networks (GANs), Large Language Models (LLMs), Mental Health Risk Prediction, Privacy-Preserving AI.

S. Moebs et al. (eds.), *Proceedings of the FIREtalk Conference - Research on FIRE! (research-on-fire 2025)*, Advances in Social Science, Education and Humanities Research 1018,
https://doi.org/10.2991/978-94-6239-705-7_10

1 Introduction

1.1 Background

Mental health disorders are a significant and growing challenge globally, affecting millions and contributing to substantial health, social, and economic burdens. According to the World Health Organization (WHO), one in eight individuals worldwide, over 970 million people, suffered from some form of mental or substance use disorder as of 2019, a figure that has likely increased in recent years due to stressors such as the COVID-19 pandemic, economic instability, and social isolation [1]. Disorders such as depression and anxiety are leading contributors to years lived with disability (YLDs), with depression alone estimated to account for 7.5% of all YLDs globally, ranking among the top causes of disability across all age groups [2].

In South Africa, the mental health crisis mirrors global trends but with additional complexities influenced by socio-economic and political challenges. Approximately 30% of South Africans will suffer from a mental health disorder during their lifetime, with conditions such as depression, anxiety, and substance use disorders being especially prevalent [3]. Moreover, the South African Stress and Health (SASH) study highlights that lifetime prevalence rates of mental health disorders in South Africa align closely with global averages, but the country faces additional barriers related to stigma, limited resources, and service accessibility [4]. These barriers exacerbate the treatment gap, where approximately 75% of South Africans with mental health disorders do not receive adequate care [5].

The burden of untreated mental health disorders extends beyond individual suffering, contributing to substantial economic losses. A study by Chisholm et al. [6] estimated that the global economic cost of mental health disorders could exceed $16 trillion by 2030 due to lost productivity and increased healthcare costs, a projection that includes significant implications for low- and middle-income countries (LMICs) like South Africa. The treatment gap in South Africa reflects a combination of underfunding, with mental health services receiving less than 5% of the total health budget, as well as a shortage of mental health professionals. For instance, South Africa has an estimated 0.31 psychiatrists and 0.32 psychologists per 100,000 population, significantly below the WHO's recommended standards [7].

Stigma remains another critical barrier, deterring individuals from seeking care and reinforcing harmful societal beliefs about mental health. According to recent studies, nearly half of South Africans perceive mental illness as a form of personal weakness, a factor that often leads to delayed treatment and exacerbates the progression of mental health disorders [8].

Artificial intelligence (AI) has increasingly been explored as a tool to support mental health care through automated screening, risk prediction, and decision support. Advances in machine learning and natural language processing have enabled the analysis of clinical records, patient narratives, and digital behavioral data to identify early indicators of mental health disorders and support timely intervention. While these approaches show promise in improving access and continuity of care, particularly in

resource-constrained settings, their effectiveness depends on ethical deployment, cultural sensitivity, and strong privacy protections.

1.2 Challenges in Mental Health AI

Despite recent advancements in artificial intelligence (AI), applying these technologies in mental health care remains challenging due to several persistent barriers. Data privacy concerns are particularly acute, given the sensitive nature of mental health records and the risk of stigma if confidentiality is breached [9]. Another challenge is that mental health datasets are often small, fragmented, and lack diversity, limiting the generalizability of predictive models. Another significant challenge is the limited interpretability of AI-driven assessments, which hinders clinical trust and practical adoption. Together, these issues underscore the need for robust, privacy-preserving, and explainable AI frameworks tailored specifically for mental health contexts.

This section will explore these challenges, beginning with a discussion of privacy and security risks associated with centralized data storage. It will then examine how data sparsity and heterogeneity impact the accuracy and reliability of machine learning models in mental health applications. It will also highlight the difficulties clinicians face in adopting AI tools due to the often opaque, "black box" nature of current deep learning models, emphasizing the critical importance of interpretability and transparency in clinical decision-making.

Data Privacy Concerns

Protecting the privacy of mental health data is paramount, as breaches can result in severe social, psychological, and economic consequences for individuals [10]. Traditional centralized machine learning approaches often require aggregation of sensitive patient information into a single repository, which increases the risk of unauthorized access, data leaks, or misuse. Additionally, mental health data is uniquely vulnerable, encompassing not only medical histories but also personal narratives, behavioral patterns, and emotional states, all of which intensify the ethical burden of maintaining confidentiality [11]. Privacy concerns are a major barrier to data sharing between healthcare institutions, limiting the availability of large, diverse datasets needed to train effective AI models. These challenges necessitate the adoption of alternative strategies, such as federated learning, which allow institutions to collaboratively train AI models without transferring sensitive data across organizational boundaries.

Privacy preservation, early detection, and risk prediction are central to addressing South Africa's mental health crisis and form the core motivation for this research. Mental health data constitutes special personal information under South Africa's Protection of Personal Information Act (POPIA), requiring stringent safeguards around consent, processing, storage, and sharing. In a context where stigma, inequality, and mistrust of institutional systems already discourage help-seeking, any perception of inadequate data protection can further suppress disclosure and engagement with mental health

services [12][13]. Privacy-preserving machine learning approaches are therefore essential to enabling ethically compliant data use while fostering public trust and alignment with national regulatory frameworks.

Many mental health disorders follow a gradual trajectory, with early cognitive, behavioral, and linguistic signals preceding acute deterioration, hospitalization, or self-harm. AI-driven early detection and risk prediction provide an opportunity to identify vulnerable individuals before crises emerge, enabling preventative interventions and more efficient use of South Africa's severely constrained mental health workforce [14][15]. By embedding POPIA-aligned privacy preservation within predictive AI models, this study aims to develop a scalable framework that supports proactive, ethically governed mental health care delivery tailored to South Africa's socio-legal and resource-constrained environment.

Data Sparsity and Heterogeneity

Mental health datasets often suffer from issues of sparsity and heterogeneity, posing significant challenges to developing accurate and generalizable AI models. Sparse datasets are common because mental health diagnoses typically require subjective assessments and extensive clinical interviews, resulting in limited structured data availability [15]. The highly individualized nature of mental health conditions also leads to heterogeneous data distributions, with wide variability across demographic groups, cultures, and clinical contexts [16]. This diversity complicates the development of models that can reliably generalize across populations without overfitting to narrow subsets of data. Traditional machine learning approaches, which rely heavily on large, homogeneous datasets, often perform poorly under such conditions, leading to biased or inaccurate predictions. Therefore, enhancing data diversity through techniques like synthetic data generation, combined with robust federated learning frameworks, is essential to address these limitations and ensure equitable AI-based mental health care.

Lack of Interpretability in AI Models

A significant barrier to the adoption of AI in mental health care is the limited interpretability of most machine learning models, particularly deep learning architectures. These models often operate as "black boxes," producing predictions without providing understandable reasoning or explanations, which creates challenges for clinicians who must justify decisions in sensitive therapeutic contexts [17]. In mental health settings, where nuanced clinical judgment and patient trust are paramount, the inability to explain AI outputs can undermine both clinician confidence and patient acceptance of AI-driven interventions [18]. Moreover, a lack of transparency raises ethical concerns regarding bias, accountability, and fairness in mental health diagnosis and treatment. Addressing these issues demands the integration of explainable AI (XAI) techniques that

can produce clinically interpretable models, ensuring that AI tools serve as trustworthy decision-support systems rather than opaque authorities.

Motivation and Objectives

The integration of artificial intelligence into mental health care offers significant potential to improve early detection, risk prediction, and continuity of care, yet its adoption remains constrained by concerns around data privacy, data sparsity, and limited model interpretability. These challenges are particularly pronounced in mental health contexts, where sensitive patient data, regulatory requirements, and resource limitations necessitate privacy-preserving and scalable solutions. This study is motivated by the need to address these barriers through the development of a novel hybrid framework that integrates federated learning, generative models, and large language models. The primary objectives of this research are to design a privacy-preserving AI architecture that enables decentralized learning without sharing raw mental health data; to enhance model robustness and generalizability through synthetic data augmentation; and to support clinically meaningful risk prediction and early detection using explainable, language-based analysis. By achieving these objectives, the study aims to contribute a scalable and ethically grounded AI framework suitable for deployment in resource-constrained mental health systems.

2 Related Work

Advances in artificial intelligence (AI) have spurred a growing body of research aimed at addressing privacy, data scarcity, and interpretability challenges in health care, and more recently, in mental health applications. This section reviews relevant studies across three key areas, which are the use of federated learning (FL) to enable privacy-preserving model training, the application of generative adversarial networks (GANs) for synthetic data generation, and the emergence of large language models (LLMs) for improving model explainability and decision support. Particular attention is given to the strengths and limitations of existing approaches, highlighting why a hybrid framework is needed to overcome persistent gaps. This section establishes the context and justification for the introduction of the proposed FL-GAN-LLM framework, by critically analyzing previous work in each of these domains.

2.1 Federated Learning in Healthcare

Overview and Key Benefits

Federated learning (FL) represents a transformative shift from traditional centralized machine learning approaches by enabling collaborative model training without the need to exchange raw data [22]. In FL, decentralized data sources, such as hospitals, clinics, and individual devices, train local models independently, and only the updated model

parameters are aggregated to form a global model [20]. This design inherently protects sensitive patient information and aligns with stringent privacy regulations like the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and in the South African context the Protection of Personal Information Act (POPIA).

In mental health contexts, where data sensitivity is even more pronounced due to stigma and confidentiality concerns, FL offers a crucial privacy-preserving advantage [23]. Furthermore, FL can facilitate the development of models that are more representative of diverse populations, as data from multiple geographical, social, and economic backgrounds can contribute to the learning process without centralized collection [24]. This decentralized capability is particularly valuable in global mental health initiatives, where resource disparities exist between regions.

Challenges Specific to Mental Health Applications

Despite its advantages, applying FL to mental health presents distinct challenges. One major issue is the heterogeneity of mental health data across different institutions. Variations in diagnostic criteria, clinical practices, language usage in therapy notes, and demographic distributions lead to non-independent and identically distributed (non-IID) data, which can significantly degrade model performance in federated settings [25].

Moreover, mental health datasets are often characterized by small sample sizes and high sparsity, especially for rare conditions or specific populations such as adolescents or minorities [15]. These characteristics exacerbate the already complex optimization processes in FL, leading to slow convergence or biased global models. Additionally, the interpretability of federated models remains a concern; in clinical mental health applications, black-box models without transparent reasoning pathways are less likely to gain clinician trust or regulatory approval [18].

These limitations underscore the need for complementary techniques, such as synthetic data augmentation and enhanced explainability mechanisms, which can be integrated into FL architectures to better suit the nuanced demands of mental health care.

2.2 GANs for Data Augmentation

Overview and Key Benefits

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [26], are a class of deep learning models that consist of two neural networks, the generator and the discriminator, that are trained in opposition to each other. Through this adversarial process, GANs are capable of generating highly realistic synthetic data that closely resembles the original dataset. In healthcare, and particularly in mental health domains,

GANs have emerged as a powerful tool to address data scarcity and imbalance issues [21].

The ability of GANs to generate synthetic mental health records, therapy session notes, or diagnostic images enables augmentation of small and imbalanced datasets without violating patient privacy [27]. This is particularly critical in mental health, where obtaining large, annotated datasets is often challenging due to ethical, regulatory, and logistical constraints [28]. By expanding the diversity of training data, GAN-augmented datasets can improve model robustness, reduce overfitting, and enable better generalization to unseen cases.

Challenges Specific to Mental Health Applications

Despite their potential, applying GANs in mental health settings presents unique challenges. First, the quality of synthetic data generated for mental health tasks, especially those involving free-text therapy notes or subjective assessments, is difficult to validate objectively [29]. Unlike structured data or imaging, mental health data often involve complex, nuanced emotional or cognitive constructs that are harder to replicate authentically.

GANs are also known to suffer from instability during training, mode collapse (where the generator produces limited varieties of outputs), and sensitivity to hyperparameters [30]. These challenges are amplified when working with small or noisy mental health datasets. Furthermore, there are ethical concerns about the use of synthetic mental health data, especially regarding the risk of inadvertently generating synthetic samples that resemble real patients too closely, thereby reintroducing privacy risks [31].

Therefore, while GANs offer a promising solution for data augmentation in mental health AI, careful methodological design, rigorous validation, and ethical oversight are essential to their successful application.

Use Cases in Medical AI

GANs have been successfully deployed across various domains of medical AI, demonstrating their versatility beyond traditional image synthesis. In radiology, GANs have been used to augment medical imaging datasets, such as MRI and CT scans, enabling improved performance in tumor detection and segmentation tasks [32]. In pathology, GAN-generated histopathology slides assist in developing models for cancer diagnosis where real annotated images are scarce [33].

In the context of mental health, use cases are emerging where GANs synthesize anonymized therapy session texts, psychiatric evaluation reports, and voice data from therapy sessions to support the development of natural language processing (NLP) models for depression and anxiety screening. Additionally, GANs have been proposed to simulate patient trajectories in mental health electronic health records, which can support predictive modeling and resource allocation planning in psychiatric services [21].

These use cases highlight the expanding role of GANs as a vital component in addressing data scarcity, privacy concerns, and model generalization challenges in healthcare AI.

Strengths and Limitations in Generating Synthetic Healthcare Data

GANs offer several strengths for synthetic data generation in healthcare. They can produce high-fidelity data that maintains complex relationships between variables, which is critical for realistic clinical modeling [34]. GANs are also capable of generating diverse datasets, which helps in mitigating model biases and improving generalization to unseen clinical scenarios.

However, GAN-generated data in healthcare settings is not without limitations. One major challenge is the potential for mode collapse, where generated samples lack variability, limiting their usefulness for training robust AI models [26]. Moreover, although synthetic, GAN outputs may inadvertently memorize and reveal sensitive patient information if not properly regularized [31]. This raises significant privacy risks, undermining one of the key motivations for using synthetic data.

Evaluating the realism and clinical validity of GAN-generated healthcare data is inherently challenging, particularly in mental health domains where subjective assessments play a central role [29]. As a result, reliance on GANs requires careful validation, expert clinical review, and integration of privacy-preserving techniques to ensure ethical and effective deployment.

2.3 Large Language Models for Mental Health

Use Cases in Medical AI

Large Language Models (LLMs) such as BERT, BioBERT, and GPT-based architectures have become foundational tools in clinical natural language processing (NLP). In healthcare, LLMs are widely used for tasks like summarizing clinical notes, extracting medical entities, supporting diagnostic decisions, and generating patient discharge summaries [35][36]. Their ability to process and contextualize unstructured clinical text has made them indispensable for making sense of narrative-heavy medical data, especially in electronic health records (EHRs).

In mental health, LLMs have been applied to assess sentiment, identify linguistic markers of psychiatric risk, and automate the analysis of therapy transcripts and patient conversations [37]. For example, ClinicalBERT, a BERT model fine-tuned on clinical notes, has shown effectiveness in predicting suicide risk and depression onset using patients' longitudinal notes [38]. These models also support early screening efforts in telepsychiatry, where language-based cues are often the only available data.

LLMs have begun to serve as explainability tools in themselves, generating narrative justifications for model outputs or producing human-readable summaries that help clinicians interpret AI-driven insights [39]. Their ability to "speak the clinician's

language" is central to building trust in AI applications within sensitive domains like mental health care.

Strengths and Limitations in Generating Synthetic Healthcare Data

LLMs also play a growing role in the generation of synthetic healthcare data. When fine-tuned on anonymized clinical corpora, they can produce realistic and structurally accurate clinical text, such as synthetic progress notes, mental health intake forms, or doctor-patient dialogues [40]. These synthetic texts can be used to augment training datasets, support simulation-based education, or develop and validate NLP algorithms where real data access is restricted.

Among their strengths, LLMs preserve long-range textual coherence and generate contextually relevant content, making them ideal for mimicking the flow of psychiatric evaluations or narrative therapy sessions. Additionally, text generated by LLMs tends to be more semantically meaningful than traditional text augmentation techniques like word shuffling or synonym replacement.

However, there are important limitations. LLMs can suffer from “hallucinations”, producing content that is fluent but factually incorrect or clinically implausible [41]. In high-stakes environments like mental health, where small language shifts may alter diagnostic interpretations, this poses a serious risk. Furthermore, there remains an open concern about the potential for these models to memorize and inadvertently reproduce fragments of their training data, which can reintroduce privacy vulnerabilities [42].

To be safely deployed for synthetic data generation, LLMs must undergo rigorous testing, bias auditing, and privacy-preservation assessments. When used responsibly, however, they offer a powerful complement to GANs in expanding and explaining mental health datasets in a privacy-conscious manner.

2.4 Identified Research Gaps

Current solutions often fail to holistically integrate privacy preservation, data augmentation, and interpretability in mental health applications, motivating this new framework.

While federated learning (FL), generative adversarial networks (GANs), and large language models (LLMs) each offer significant promise in healthcare AI, their isolated application often falls short when addressing the unique complexities of mental health care. FL offers privacy-preserving benefits but struggles with data heterogeneity and limited interpretability, particularly in scenarios where local mental health datasets are sparse, unstructured, or non-standardized [43]. Without additional mechanisms to enrich and harmonize these datasets, federated models may converge slowly or perform unevenly across sites.

GANs help address data scarcity by producing synthetic samples, yet their integration into real-world mental health settings is still limited due to evaluation difficulties and concerns about fidelity and safety of generated data [29]. Meanwhile, although

LLMs enhance interpretability and simulate nuanced clinical language, they can introduce hallucinated content and privacy vulnerabilities if not carefully validated [42]. Furthermore, the black-box nature of these systems continues to raise concerns around bias, clinical accountability, and ethical transparency.

A key gap in the literature is the lack of integrated frameworks that combine these three technologies in a way that leverages their strengths while compensating for their individual limitations. For example, FL alone cannot solve data sparsity; GANs can augment data but need local tuning; LLMs require extensive contextual training and safeguards. Addressing these gaps requires a unified, layered approach capable of operating across decentralized systems, enriching sparse data intelligently, and providing interpretable outputs to clinical end-users.

This observation forms the foundation for the proposed FL-GAN-LLM framework introduced in the following section, a multi-layered architecture designed specifically to meet the data, privacy, and interpretability needs of mental health AI.

Proposed Framework: Federated GANS-LLM

To effectively address the pressing challenges in mental health AI, this study proposes a novel hybrid framework that strategically integrates federated learning (FL), generative adversarial networks (GANs), and large language models (LLMs). The FL component ensures that model training occurs locally across multiple institutions, safeguarding sensitive patient data and complying with privacy regulations such as GDPR, POPIA and HIPAA [19]. However, given the sparsity and heterogeneity of mental health datasets, relying solely on FL may not achieve optimal model generalization. To mitigate this, the framework incorporates GANs to synthetically augment local datasets, enhancing diversity without compromising privacy [21]. To also address the interpretability gap inherent in deep learning systems, an LLM layer is proposed to generate human-readable explanations of model predictions, tailored specifically for clinical settings. The hybrid FL-GAN-LLM architecture thus aims to balance privacy, data richness, and explainability, positioning it as a promising blueprint for future mental health AI applications.

The proposed framework adopts a multi-layered architecture that integrates federated learning (FL), generative adversarial networks enhanced with variational autoencoders (GAN-VAEs), and large language models (LLMs) to balance privacy, data enrichment, and interpretability in mental health applications. The process is explained in the next section and the Figure 1 depicts the architecture.

Fig. 1.

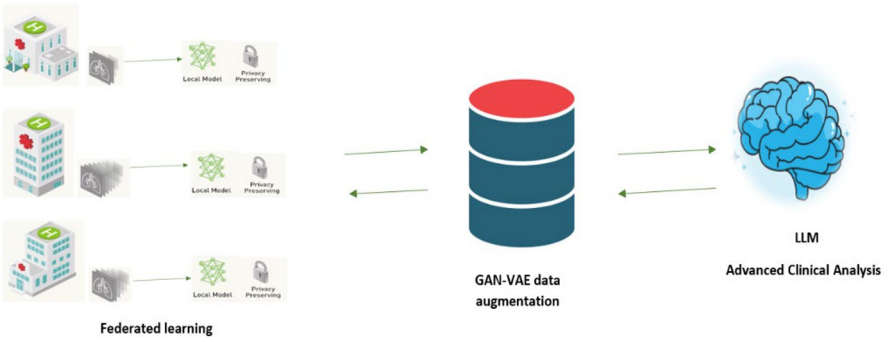

Figure 2 Architectural design

2.5 Client-Side Federated Learning Layer

At the client level, the architecture deploys federated learning to allow each healthcare institution to train its local model on sensitive patient data without directly sharing that data. Each participating hospital employs a lightweight embedding model as part of a split learning process, extracting key features from clinical data such as electronic health records and patient notes. Instead of transmitting raw data or intermediate activations, institutions compute model updates (e.g., gradients or weight changes), encrypt these updates, and then transmit them securely to a central server.

To maintain strict confidentiality, encryption is applied using secure aggregation protocols based on homomorphic encryption. For example, the Paillier cryptosystem offers additive homomorphic properties that allow each client to encrypt its model update in such a way that the central server can aggregate the encrypted values without ever decrypting or viewing individual updates. This method was operationalized in Bonawitz et al.'s (2017) secure aggregation protocol, which ensures that only aggregated updates are revealed to the server, thus preserving the privacy of each client's data. Other works, such as CryptoNets [20], have demonstrated the feasibility of applying homomorphic encryption to neural networks, further validating its use in federated mental health applications.

This approach ensures that while local models benefit from collective training, patient-level data confidentiality is preserved throughout the process, a critical requirement in sensitive mental health contexts.

2.6 GAN-Based Data Augmentation Layer

Once the encrypted updates reach the central server, a data augmentation module takes center stage. Encrypted model updates from all clients are first aggregated using secure aggregation protocols. These protocols allow the server to compute collective updates

without exposing individual contributions. The resulting aggregated values are then decrypted, yielding privacy-preserving intermediate activations. These activations form the input to a hybrid GAN-VAE model, which generates high-quality synthetic data to mitigate issues of data sparsity and heterogeneity.

The integration of variational autoencoders (VAEs) with GANs is particularly important in this setting. While GANs are highly effective at producing realistic synthetic data, they often face instability issues such as mode collapse. Incorporating VAEs helps stabilize training, enhances diversity, and improves the realism of synthetic clinical data [44]. This hybrid architecture has been shown to enrich datasets and improve downstream model robustness [45].

By leveraging this GAN-VAE module, the system produces enriched datasets that represent the diversity and complexity of clinical information more faithfully, making subsequent analysis by language models more accurate and clinically meaningful.

2.7 ClinicalBERT-Driven Risk Assessment Layer

The enriched dataset generated by the GAN-VAE module is processed by ClinicalBERT, a variant of the Bidirectional Encoder Representations from Transformers (BERT) model that has been specifically adapted for clinical text. ClinicalBERT is designed to capture the nuances of clinical language and extract meaningful patterns from complex medical narratives, making it well-suited for applications in mental health care. Huang, Altosaar, and Ranganath (2019) demonstrated that ClinicalBERT consistently outperforms traditional language models when applied to clinical notes, underlining its superior contextual understanding [38].

Within this framework, ClinicalBERT performs advanced contextual analysis, risk prediction, and sentiment analysis, extracting key indicators of potential mental health risks such as relapse likelihood or hospitalization risk. The insights and predictions generated by ClinicalBERT are then fed back into the client-side federated models as gradient updates. This feedback loop enables local models to continuously refine their parameters, thereby improving predictive accuracy over time.

Finally, to enhance the transparency of model outputs, the framework integrates interpretability mechanisms. Model-agnostic explanation methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) provide feature importance scores that illustrate which clinical input elements most strongly influence each prediction [46][47]. In addition, the attention mechanisms inherent in ClinicalBERT can be extracted and visualized, offering insight into how specific words or phrases within clinical notes drive decision-making. This dual interpretability strategy combines quantitative feature contribution analysis with qualitative, context-specific explanations, ensuring accountability and building clinician trust in the AI system.

3 Methodological Considerations

The proposed framework is at a conceptual stage and requires careful planning for evaluation, validation, and eventual deployment. This section outlines the methodological approach that will guide the study. It begins by discussing the planned evaluation strategies, including the use of synthetic and benchmark datasets to test model feasibility and performance. Next, it addresses ethical and privacy considerations, particularly compliance with relevant regulations and data governance frameworks, which are critical given the sensitivity of mental health data. It also considers potential deployment challenges, such as computational overhead, communication efficiency, and model drift, that may arise when scaling the system across multiple institutions. By integrating these methodological considerations, the study aims to ensure both the technical robustness and ethical integrity of the proposed FL-GAN-LLM framework.

3.1 Planned Evaluation Strategies

The evaluation of the proposed FL-GAN-LLM framework will proceed in multiple stages to establish its feasibility, technical performance, and potential clinical relevance. Since the framework has not yet been tested in real-world settings, the initial phase will rely on synthetic and benchmark datasets. Publicly available resources such as the MIMIC-III clinical database [48] and other de-identified text corpora will serve as a starting point for testing the individual components of the system. Synthetic datasets generated through GAN-VAE models will be used to supplement these resources, enabling the assessment of the framework's ability to address data sparsity and heterogeneity challenges.

Model performance will be evaluated using established metrics. For the federated learning module, metrics such as model convergence rates, accuracy, and F1-scores will be assessed across distributed clients [20]. The GAN-VAE component will be evaluated using fidelity and diversity measures, including the Fréchet Inception Distance (FID) and domain-specific validation by clinical experts to assess the plausibility of synthetic data [49]. For the ClinicalBERT-driven LLM module, evaluation will focus on its predictive accuracy, interpretability, and ability to detect risk markers. Metrics such as Area Under the Receiver Operating Characteristic Curve (AUROC), precision-recall curves, and attention-weight analyses will be applied [50].

Finally, to ensure the framework's clinical utility, interpretability and trustworthiness will be key evaluation criteria. Model-agnostic explanation tools such as SHAP and LIME will be applied to validate that those predictions are not only accurate but also explainable in ways meaningful to clinicians [47][46]. Through this multi-layered evaluation strategy, the study aims to demonstrate the framework's capacity to achieve robust predictive performance while maintaining transparency and safeguarding patient privacy.

3.2 Ethical and Privacy Considerations

The framework will adhere to data protection regulations such as POPIA and GDPR, incorporating principles of ethical AI design. Given the highly sensitive nature of mental health data, ethical and privacy safeguards are central to the design and deployment of the proposed framework. Data used in this study will be de-identified wherever possible, and strict governance protocols will be observed to ensure compliance with relevant privacy regulations, including the General Data Protection Regulation (GDPR) in Europe and the Protection of Personal Information Act (POPIA) in South Africa. These frameworks emphasize data minimization, purpose limitation, and accountability, all of which will be embedded in the framework's data-handling procedures [16].

The use of federated learning mitigates many privacy risks by ensuring that raw patient data remain within institutional boundaries, eliminating the need for centralized data pooling. In addition, model updates will be protected through secure aggregation protocols employing homomorphic encryption, ensuring that individual contributions cannot be reconstructed by the central server [51]. The incorporation of synthetic data through the GAN-VAE module adds another privacy-preserving layer by reducing reliance on small, sensitive datasets while still maintaining clinical relevance.

Beyond legal compliance, the framework will also adhere to ethical principles for AI in healthcare, including fairness, transparency, and accountability [52]. Mechanisms such as SHAP and LIME will provide explainability, reducing the "black box" effect and supporting clinician trust in AI-assisted decision-making. Regular auditing will be necessary to identify and mitigate bias, particularly in datasets that underrepresent vulnerable populations, as bias in mental health AI can exacerbate existing disparities [29].

Finally, ethical oversight will be sought through institutional review boards (IRBs) and interdisciplinary advisory panels, ensuring that the framework not only meets technical standards but also respects patient dignity, autonomy, and trust. By integrating privacy-preserving design with robust ethical governance, this study aims to set a precedent for responsible AI development in mental health.

3.3 Potential Deployment Challenges

Challenges include ensuring computational efficiency, managing communication overhead between clients and server, and addressing model drift over time. While the proposed FL-GAN-LLM framework offers strong potential for advancing privacy-preserving and interpretable AI in mental health, several challenges must be anticipated in moving from conceptual design to practical deployment.

First, computational resource constraints represent a significant barrier. Federated learning and GAN-VAE training are computationally intensive, requiring substantial processing power and memory across both client institutions and central servers. Many mental health facilities, particularly in low-resource settings, may lack the necessary infrastructure to participate fully in such a system [53]. Ensuring scalability will

therefore require optimization strategies, such as model compression or lightweight architectures.

Second, communication overhead poses difficulties in federated environments. The iterative exchange of encrypted model updates between clients and the server may result in high network bandwidth demands, particularly when scaled to multiple institutions. This challenge has been well-documented in federated systems and can be addressed through techniques such as update sparsification, quantization, or asynchronous communication protocols [20].

Third, model drift and heterogeneity are persistent concerns. Differences in patient demographics, diagnostic practices, and clinical documentation styles across institutions can result in non-IID (non-independent and identically distributed) data, reducing the global model's generalizability and slowing convergence [43]. Continuous monitoring, adaptive weighting of client contributions, and the use of synthetic augmentation are potential strategies to mitigate these issues.

Finally, trust and adoption barriers must not be underestimated. Even with privacy safeguards and explainability tools, clinicians may remain hesitant to rely on AI for high-stakes mental health decisions. Addressing this requires not only technical reliability but also co-design with clinicians, transparent communication of model limitations, and training initiatives to integrate AI into clinical workflows [18].

Anticipating and addressing these deployment challenges is crucial to ensure that the framework moves beyond theoretical promise to become a practical and trusted tool in real-world mental health care.

3.4 Anticipated Contributions and Impact

Technical Contributions.

The framework offers a novel integration of FL, GANs, and LLMs for mental health applications, advancing the state of privacy-preserving AI systems. This study will advance the field of privacy-preserving AI by introducing a multi-layered architecture that integrates federated learning, GAN-VAE data augmentation, and ClinicalBERT-driven risk assessment. Specifically, it contributes:

- A novel integration of FL and GAN-VAE models that enables both privacy preservation and synthetic data generation, addressing the twin challenges of confidentiality and data sparsity.
- The application of ClinicalBERT within a federated setting, offering contextualized analysis of clinical narratives for mental health risk detection with improved interpretability.
- An iterative feedback loop where insights from the LLM inform local federated models, thereby improving predictive accuracy over time.
- The incorporation of interpretability mechanisms such as SHAP, LIME, and attention visualization to provide clinicians with transparent, actionable insights.

These technical innovations collectively demonstrate how cutting-edge AI methods can be combined into a coherent, scalable framework tailored for sensitive domains like mental health.

3.5 Clinical and Societal Impact

By enhancing early risk detection while preserving patient privacy, this work has the potential to significantly bridge mental health treatment gaps, particularly in underserved communities. Beyond technical innovation, the framework aspires to create meaningful clinical and social value. By enabling early risk prediction and detection of mental health issues, the system can help clinicians intervene proactively, potentially reducing relapse rates and hospitalizations. Its privacy-preserving design ensures that sensitive patient information remains protected, thereby fostering trust among patients, clinicians, and institutions.

Addressing data sparsity and heterogeneity, the framework has the potential to improve equity in mental health care, particularly in low-resource contexts where data are limited and disparities in access are severe [54]. The scalable nature of federated learning means that even institutions with limited infrastructure can contribute to and benefit from a global model without compromising data ownership or security.

The proposed architecture is not only a technological advancement but also a step toward bridging the global mental health treatment gap, particularly in under-resourced regions such as South Africa. If successfully developed and deployed, this framework could serve as a blueprint for privacy-conscious, scalable AI applications across other areas of healthcare as well.

4 Conclusion and Future Work

This paper has presented a conceptual framework, the Federated GAN-LLM with VAE architecture, designed to advance privacy-preserving artificial intelligence for mental health risk prediction and early detection. By combining federated learning to safeguard sensitive patient data, a hybrid GAN-VAE module to address data sparsity and heterogeneity, and ClinicalBERT to provide contextualized risk assessment with interpretability features, the framework aims to overcome the key barriers that currently limit the use of AI in mental health care.

The framework's novelty lies in its multi-layered integration: a secure client-side learning process that prevents raw data sharing; a centralized synthetic data augmentation strategy to strengthen global model performance; and a large language model equipped with interpretability mechanisms to ensure transparency and trust. Together, these elements position the framework as a scalable, ethically grounded solution for bridging gaps in mental health services, particularly in under-resourced contexts such as South Africa.

As this study represents an early-stage PhD proposal, no experimental results have yet been obtained. Future work will focus on developing a prototype of the framework, beginning with feasibility testing on benchmark and synthetic datasets, followed by pilot studies with clinical partners. Emphasis will also be placed on refining interpretability methods, conducting fairness audits to reduce algorithmic bias, and ensuring regulatory compliance across multiple jurisdictions. Ultimately, the long-term goal is

to move from conceptual development to real-world validation, demonstrating the practical impact of privacy-preserving, explainable AI in improving mental health outcomes.

References

1. "The (Mental Health) State of the World: 2022 WHO Report." Accessed: Mar. 15, 2025. [Online]. Available: https://www.neurocaregroup.com/news-insights/the-mental-health-state-of-the-world-summary-of-2022-who-mental-health-report
2. T. Vos et al., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," Lancet, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020, doi: 10.1016/S0140-6736(20)30925-9.
3. A. Herman, D. Stein, S. Seedat, S. Heeringa, H. Moomal, and D. R. Williams, "The South African Stress and Health (SASH) study: 12-month and lifetime prevalence of common mental disorders - PubMed." Accessed: Mar. 15, 2025. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19588796/
4. P. Monnapula-Mazabane and I. Petersen, "Mental health stigma experiences among caregivers and service users in South Africa: a qualitative investigation," Curr. Psychol., vol. 42, no. 11, pp. 9427–9439, Apr. 2023, doi: 10.1007/s12144-021-02236-y.
5. W. H. O. (WHO), "Mental health atlas," 2021.
6. D. Chisholm et al., "Scaling-up treatment of depression and anxiety: A global return on investment analysis," The Lancet Psychiatry, vol. 3, no. 5, pp. 415–424, May 2016, doi: 10.1016/S2215-0366(16)30024-4.
7. S. Docrat, D. Besada, S. Cleary, E. Daviaud, and C. Lund, "Mental health system costs, resources and constraints in South Africa: a national survey," Health Policy Plan., vol. 34, no. 9, pp. 706–719, Nov. 2019, doi: 10.1093/HEAPOL/CZZ085.
8. C. Lund et al., "Social determinants of mental disorders and the Sustainable Development Goals: a systematic review of reviews.," The lancet. Psychiatry, vol. 5, no. 4, pp. 357–369, Apr. 2018, doi: 10.1016/S2215-0366(18)30060-9.
9. E. Watson, S. Fletcher-Watson, and E. J. Kirkham, "Views on sharing mental health data for research purposes: qualitative analysis of interviews with people with mental illness," BMC Med. Ethics, vol. 24, no. 1, pp. 1–12, Dec. 2023, doi: 10.1186/s12910-023-00961-6.
10. W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," Nature Medicine, vol. 25, no. 1. Nature Publishing Group, pp. 37–43, Jan. 01, 2019. doi: 10.1038/s41591-018-0272-7.
11. A. Mandal, T. Chakraborty, and I. Gurevych, "Towards Privacy-aware Mental Health AI Models: Advances, Challenges, and Opportunities," Feb. 2025, Accessed: May 05, 2025. [Online]. Available: https://arxiv.org/pdf/2502.00451
12. Popia, "Protection of Personal Information Act (POPI Act) - POPIA," Popia. Accessed: Apr. 26, 2025. [Online]. Available: https://popia.co.za/
13. L. C et al., "Social determinants of mental disorders and the Sustainable Development Goals: a systematic review of reviews.," The lancet. Psychiatry, vol. 5, no. 4, pp. 357–369, 2018, Accessed: Mar. 28, 2025. [Online]. Available: https://www.researchgate.net/publication/324056783_Social_determinants_of_mental_disorders_and_the_Sustainable_Development_Goals_a_systematic_review_of_reviews
14. C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting Risk of Suicide Attempts Over Time Through Machine Learning," Clin. Psychol. Sci., vol. 5, no. 3, pp. 457–469, 2017, doi: 10.1177/2167702617691560.

15. S. T. A. Shatte, D. Hutchinson, “Machine learning in mental health: A systematic scoping review of methods and applications Adrian B. R. Shatte*,” 2021.
16. S. Pati et al., “Privacy preservation for federated learning in health care,” Patterns, vol. 5, no. 7. Cell Press, p. 100974, Jul. 12, 2024. doi: 10.1016/j.patter.2024.100974.
17. F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” Feb. 2017, Accessed: May 05, 2025. [Online]. Available: https://arxiv.org/pdf/1702.08608
18. S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, “What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use,” in Proceedings of Machine Learning Research, PMLR, Oct. 2019, pp. 359–380. Accessed: May 05, 2025. [Online]. Available: https://proceedings.mlr.press/v106/tonekaboni19a.html
19. N. Rieke et al., “The future of digital health with federated learning,” npj Digit. Med., vol. 3, no. 1, pp. 1–7, Dec. 2020, doi: 10.1038/s41746-020-00323-1.
20. P. Kairouz et al., “Advances and open problems in federated learning,” Foundations and Trends in Machine Learning, vol. 14, no. 1–2. Now Publishers Inc, pp. 1–210, Jun. 23, 2021. doi: 10.1561/2200000083.
21. C. Esteban, S. L. Hyland, and G. Rätsch, “Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs,” Jun. 2017, Accessed: Apr. 22, 2025. [Online]. Available: https://arxiv.org/abs/1706.02633v2
22. Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” ACM Trans. Intell. Syst. Technol., vol. 10, no. 2, p. 19, Feb. 2019, doi: 10.1145/3298981.
23. S. Chakrabarti, “Digital psychiatry in low-and-middle-income countries: New developments and the way forward,” World J. Psychiatry, vol. 14, no. 3, pp. 350–361, 2024, doi: 10.5498/wjp.v14.i3.350.
24. M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, “Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2019, pp. 92–104. doi: 10.1007/978-3-030-11723-8_9.
25. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated Learning: Challenges, Methods, and Future Directions,” IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, May 2020, doi: 10.1109/MSP.2020.2975749.
26. I. J. Goodfellow et al., “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680. doi: 10.1007/978-3-658-40442-0_9.
27. E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, “Generating Multi-label Discrete Patient Records using Generative Adversarial Networks,” vol. 68, pp. 1–20, 2017, [Online]. Available: http://arxiv.org/abs/1703.06490
28. M. K. Baowaly, C. C. Lin, C. L. Liu, and K. T. Chen, “Synthesizing electronic health records using improved generative adversarial networks,” J. Am. Med. Informatics Assoc., vol. 26, no. 3, pp. 228–241, Mar. 2019, doi: 10.1093/JAMIA/OCY142,.
29. R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, “Synthetic data in machine learning for medicine and healthcare,” Nat. Biomed. Eng., vol. 5, no. 6, pp. 493–497, Jun. 2021, doi: 10.1038/S41551-021-00751-8,.
30. M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” Jan. 2017, Accessed: Jul. 10, 2025. [Online]. Available: https://arxiv.org/pdf/1701.07875

31. A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, “Generation and evaluation of privacy preserving synthetic health data,” Neurocomputing, vol. 416, pp. 244–255, 2020, doi: 10.1016/j.neucom.2019.12.136.
32. M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic data augmentation using GAN for improved liver lesion classification,” Proc. - Int. Symp. Biomed. Imaging, vol. 2018-April, pp. 289–293, May 2018, doi: 10.1109/ISBI.2018.8363576.
33. F. Mahmood, R. Chen, and N. J. Durr, “Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training,” IEEE Trans. Med. Imaging, vol. 37, no. 12, pp. 2572–2581, Dec. 2018, doi: 10.1109/TMI.2018.2842767,.
34. H. C. Shin et al., “Medical image synthesis for data augmentation and anonymization using generative adversarial networks,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11037 LNCS, pp. 1–11, 2018, doi: 10.1007/978-3-030-00536-8_1/TABLES/1.
35. J. Lee et al., “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.
36. A. Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” npj Digit. Med., vol. 1, no. 1, pp. 1–10, 2018, doi: 10.1038/s41746-018-0029-1.
37. I. Levkovich, “Evaluating Diagnostic Accuracy and Treatment Efficacy in Mental Health: A Comparative Analysis of Large Language Model Tools and Mental Health Professionals,” Eur. J. Investig. Heal. Psychol. Educ., vol. 15, no. 1, Jan. 2025, doi: 10.3390/ejihpe15010009.
38. K. Huang, J. Altosaar, and R. Ranganath, “ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission,” Apr. 2019, Accessed: Jul. 12, 2025. [Online]. Available: https://arxiv.org/pdf/1904.05342
39. K. Singhal et al., “Large language models encode clinical knowledge,” Nature, vol. 620, no. 7972, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.
40. E. Lehman, S. Jain, K. Pichotta, Y. Goldberg, and B. C. Wallace, “Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?,” NAACL-HLT 2021 - 2021 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf., pp. 946–959, 2021, doi: 10.18653/v1/2021.naacl-main.73.
41. P. Cruz-Gonzalez et al., Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications, vol. 55. 2025. doi: 10.1017/S0033291724003295.
42. N. Carlini et al., “Extracting training data from large language models,” Proc. 30th USENIX Secur. Symp., pp. 2633–2650, 2021.
43. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated Learning: Challenges, Methods, and Future Directions,” IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, 2020, doi: 10.1109/MSP.2020.2975749.
44. D. J. Rezende and F. Viola, “Taming VAEs,” 2018, [Online]. Available: http://arxiv.org/abs/1810.00597
45. H. Antoniou, A., Storkey, A., Edwards, “Data Augmentation Generative Adversarial Networks.,” 2017.
46. S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.
47. M. T. Ribeiro, S. Singh, and C. Guestrin, “‘why should i trust you?’ explaining the predictions of any classifier,” in NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,

Proceedings of the Demonstrations Session, Association for Computing Machinery, Aug. 2016, pp. 97–101. doi: 10.18653/v1/n16-3020.
48. A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” Sci. Data, vol. 3, May 2016, doi: 10.1038/SDATA.2016.35,.
49. M. K. Baowaly, C. C. Lin, C. L. Liu, and K. T. Chen, “Synthesizing electronic health records using improved generative adversarial networks,” J. Am. Med. Informatics Assoc., vol. 26, no. 3, pp. 228–241, Mar. 2019, doi: 10.1093/JAMIA/OCY142,.
50. A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, “Ensuring fairness in machine learning to advance health equity,” Ann. Intern. Med., vol. 169, no. 12, pp. 866–872, Dec. 2018, doi: 10.7326/M18-1990.
51. Keith Bonawitz and Vladimir Ivanov and Ben Kreuter and Antonio Marcedone and H. Brendan McMahan and Sarvar Patel and Daniel Ramage and Aaron Segal and Karn Seth, “Practical Secure Aggregation for Privacy Preserving Machine Learning,” vol. 5, no. 1, pp. 7–22, 2019, [Online]. Available: https://eprint.iacr.org/2017/281
52. J. Whittlestone, A. Alexandrova, R. Nyrup, and S. Cave, “The role and limits of principles in AI ethics: Towards a focus on tensions,” in AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, Inc, Jan. 2019, pp. 195–200. doi: 10.1145/3306618.3314289.
53. N. Rieke et al., “The future of digital health with federated learning,” npj Digit. Med., vol. 3, no. 1, pp. 1–7, Sep. 2020, doi: 10.1038/s41746-020-00323-1.
54. A. Mandal, P. K. Adhikary, H. Arnaout, I. Gurevych, and T. Chakraborty, “A Comprehensive Review of Datasets for Clinical Mental Health AI Systems,” Aug. 2025, Accessed: Sep. 02, 2025. [Online]. Available: https://www.arxiv.org/pdf/2508.09809.