Federated Explainable Mental Health Analytics (FEMHA): A Sustainable Framework for SDG-Aligned Risk Prediction and Emerging Challenges

Dhruv Kaushik
UC Santa Barbara
California, USA
dhkaushik20j@gmail.com

Ramu Yadavalli
Department of C.S.E.
Shri Vishnu Engineering College for Women
Bhimavaram, Andhra Pradesh, India
yramumail@gmail.com

2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON) | 979-8-3315-3482-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/I2ITCON65200.2025.11210548

Abstract - **Mental health challenges among students present a critical global issue, significantly influencing academic outcomes, personal development, and long-term societal participation. Artificial Intelligence (AI)-driven analytics offer transformative potential for early detection and intervention, improving mental health outcomes in educational contexts. This paper presents a comprehensive synthesis of twenty-five peer-reviewed studies focusing on AI techniques for student mental health risk prediction, identifying key trends, methodological approaches, emerging challenges, and research gaps. To address these gaps, we propose the Federated Explainable Mental Health Analytics (FEMHA) Framework, a sustainable, privacy-preserving, and explainable architecture. FEMHA integrates multimodal data collection, federated learning, explainable decision-making modules, and Sustainable Development Goals (SDG)-aligned outcome optimization. The framework aims to mitigate dataset privacy risks, enhance model transparency, enable equitable deployment across diverse student populations, and support SDG-3 (Good Health and Well-being), SDG-4 (Quality Education), and SDG-10 (Reduced Inequalities). Through thematic analysis, architectural modelling, and proposed best practices, this work contributes a novel, ethical, and sustainable direction for advancing AI-driven mental health analytics in student populations.**

Keywords— Mental Health Analytics, Student Depression Prediction, Federated Learning, Explainable AI, Sustainable Development Goals, Multimodal Data Fusion, Privacy-Preserving Machine Learning, Ethical AI, Predictive Analytics, Healthcare AI.

I. Introduction

Mental health issues such as depression, anxiety, and stress-related disorders are rising at an alarming rate, particularly among students, posing serious implications for personal well-being, academic success, and societal development [1]. Artificial Intelligence (AI) has become a promising tool in this space, offering new ways to detect, monitor, and support mental health through scalable and data-driven interventions [2]. By analyzing diverse health data, AI systems can identify early warning signs and provide personalized support, which is especially important in light of the global shortage of mental health professionals and persistent stigma around seeking help [3]. Healthcare analytics powered by AI not only enhances predictive capabilities and treatment planning but also supports the SDGs, including SDG-3 (Good Health and Well-Being) and SDG-4 (Quality Education), by fostering healthier, more equitable learning environments [4].

The primary scope of this review paper is to explore how AI-driven techniques can be effectively applied in student mental health care, with a focus on depression detection and management. We aim to review the current state-of-the-art models, analyze their performance and limitations, and propose a future roadmap. The motivation stems from the critical need for intelligent, ethical, and accessible solutions that can empower early interventions, foster resilience among youth, and contribute meaningfully to sustainable global development

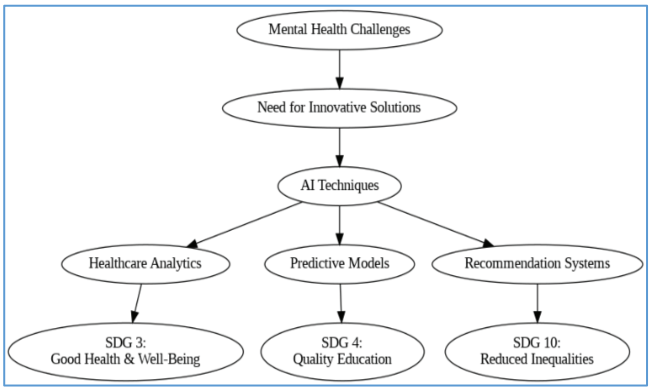

Fig. 1. Mapping of mental health challenges to SDGs

Figure-1, visualizes the mapping of mental health challenges to AI-driven techniques and their contribution to SDGs. AI applications such as healthcare analytics, predictive modelling, and recommendation systems are shown to align with key SDG objectives, reinforcing the critical role of intelligent systems in promoting mental health and global well-being.

Through this study, we aim to bridge the gap between student mental health challenges, AI-driven prediction models, and the broader goals of sustainable development. By proposing the Federated Explainable Mental Health Analytics (FEMHA) framework, this work offers an integrated perspective for researchers, clinicians, and policymakers seeking scalable, ethical, and SDG-aligned solutions in mental health analytics.

II. Research Objectives and Scope

This study is based on a systematic selection of 25 peer-reviewed articles focusing on the application of AI techniques to address mental health challenges and align with SDG's. The sources primarily include indexed journals and conferences.

The article selection process was guided by specific keywords, including "Mental Health AND Artificial Intelligence," "Depression Detection using AI," "Healthcare Analytics for Mental Health," and "AI and SDG-3 Well-being." Studies were included if they were published between

2018 and 2025, proposed or applied AI models in the context of mental health, or explicitly connected mental health interventions to Sustainable Development Goal (SDG) targets. Exclusion criteria ruled out non-peer-reviewed articles, studies focused purely on physical health or non-AI mental health therapies, and papers lacking a direct link to SDG objectives or healthcare analytics. To maintain a clear and systematic approach, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed throughout the selection process.

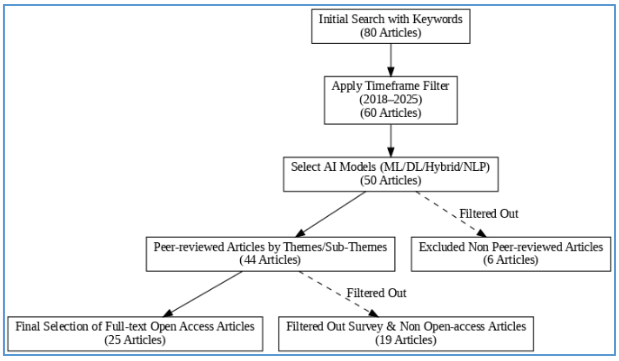
Initial Search with Keywords
(80 Articles)
Apply Timeframe Filter
(2018–2025)
(60 Articles)
Select AI Models (ML/DL/Hybrid/NLP)
(50 Articles)
Filtered Out
Peer-reviewed Articles by Themes/Sub-Themes
(44 Articles)
Excluded Non Peer-reviewed Articles
(6 Articles)
Filtered Out
Final Selection of Full-text Open Access Articles
(25 Articles)
Filtered Out Survey & Non Open-access Articles
(19 Articles)

Fig. 2. PRISMA Flow Diagram for Article Selection

Figure-2, illustrates the PRISMA flow to represent the process of identifying, screening, and including the 25 articles selected for this systematic review. The process begins with an initial search (80 articles), applying a timeframe filter (2018–2025), selecting articles focused on AI models, such as Machine Learning (ML), Deep Learning (DL), Hybrid Models, Natural Language Processing (NLP) techniques etc, and then systematically filtering out non-peer-reviewed, survey-only, and non-open-access articles. The final selection resulted in 25 relevant articles categorized by themes and sub-themes.

To guide this literature study and the development of the proposed FEMHA framework, the following key research questions were explored:

RQ1: What type of mental health concerns, such as depression, anxiety, and stress, - are frequently addressed in AI-based research aiming on student populations?

RQ2: What kinds of artificial intelligence techniques are being used, and their performance?

RQ3: What types of data sources (text, audio, clinical records, or sensor data) are being used in these studies?

RQ4: Where do current studies fall short, and how do they contribute to global development goals like SDG 3 (Health), SDG 4 (Education), and SDG 10 (Equity)?

III. Synthesis Analytical Thematic Insights of AI-Driven Mental-Health Analytics

A. *Research Problems in Mental Health Analytics:*

Mental health analytics, particularly for students, has emerged as a critical area of research due to the increasing prevalence of psychological issues such as depression, anxiety, and stress. Across the 25 selected studies, the primary research problems addressed are grouped into specific domains based on the type of mental health condition or psychological challenge targeted.

The most dominant problem domain was depression prediction and analysis, appearing in approximately 68% of the reviewed articles [1]-[5], [8], [10], [11], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [25]. Depression detection efforts leveraged traditional machine learning algorithms such as Support Vector Machines (SVM), Random Forest (RF), and also adopted DL models like Convolutional Neural Networks (CNN) and Transformer architectures [5], [15].

Anxiety prediction was the next significant focus area, addressed independently or alongside depression [4], [5], [6], [7], [8], [9], [18], [19], [23]. Researchers recognized that early identification of anxiety symptoms among students could significantly enhance intervention strategies, especially during transitional academic periods or crisis scenarios such as the COVID-19 pandemic [4].

Stress detection formed another essential subdomain. A few studies directly tackled stress prediction using physiological signals, behavioral features, or self-reported survey data [5], [17]. Although stress-related works were fewer compared to depression and anxiety, they highlighted the necessity of including stress management in overall student mental health frameworks.

A minority of studies expanded their scope to general mental health risk prediction, combining depression, anxiety, and stress into composite models for holistic well-being assessments [2], [7], [18], [23].

It is evident that while depression remains the predominant research problem, there is a growing tendency toward multi-condition predictive modeling, recognizing the comorbidity and overlapping symptoms that frequently occur among young populations.

B. *Data Sources and Modalities:*

The success of AI techniques in mental health analytics depends significantly on the nature and quality of the data sources used. A review of the 25 selected studies reveals the diversity of data modalities adopted for predicting mental health conditions such as depression, anxiety, and stress.

1) *Text Data:*

Text-based sources were the most prevalent, utilized in approximately 60% of the reviewed articles [2], [5], [8], [10], [15], [18], [22]. These included self-reported survey data, social media posts, clinical narratives, and mobile app logs. Sentiment analysis, topic modeling, and NLP techniques were widely applied to extract linguistic patterns indicative of mental health states [15], [22].

2) *Voice Data:*

A few studies incorporated audio or voice data to analyze speech patterns, prosody, and acoustic features for depression and anxiety prediction [2], [6], [11]. Voice data offers non-invasive, real-time insights into emotional states but remains under-explored due to technical and privacy challenges.

3) *Image Data:*

Limited but growing research explored the use of facial image data and expression analysis [2], [5]. Studies utilized CNNs to interpret micro-expressions or eye movements that correlate with depressive or anxious behavior.

4) *Clinical and Sensor Data:*

Several articles employed clinical notes, wearable sensor data, smartphone metadata, and physiological measurements [4], [8], [11], [16], [19]. These modalities enabled passive monitoring of health states, capturing subtle behavioural and physiological changes indicative of psychological distress.

Overall, text data remains dominant, but there is a growing trend towards multi-modal approaches combining textual, acoustic, and sensor information to improve prediction accuracy and robustness [2], [11].

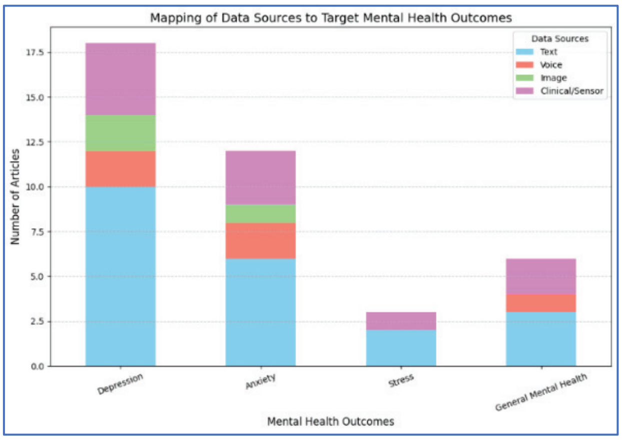

Fig. 3. Mapping of Data Sources to Target Mental Health Outcomes

This stacked bar chart in Figure-3, illustrates the mapping between different types of data sources (text, voice, image, clinical/sensor data) and the primary mental health outcomes they aim to predict (depression, anxiety, stress, general mental health risk). It highlights the dominance of text data for depression prediction. Voice and sensor data are gaining importance for multi-condition assessments. Image data remains limited but shows potential for specific anxiety and depression prediction cases.

The shift toward multi-modal data fusion reflects an effort to build more comprehensive, realistic mental health monitoring systems, moving beyond single-modality predictions. Future directions emphasize real-time, passive data collection balanced against privacy-preserving analytics, contributing directly to SDG-3 and SDG-10.

C. *Prediction Objectives:*

The reviewed studies across mental health analytics reveal three major categories of prediction objectives: diagnosis, risk prediction, and treatment outcome forecasting. Each objective area leverages specific types of AI models, aligned with the nature of the target problem and the available datasets.

1) *Diagnosis Prediction:*

Diagnosis-focused models aim to identify the presence or absence of mental health conditions such as depression, anxiety, or stress. Techniques such as SVM, RF, and CNN-based DL models are widely adopted [1], [5], [8], [15]. Textual features from surveys or social media posts are predominantly used for this purpose [15], [22].

2) *Risk Prediction:*

Risk prediction models focus on estimating the likelihood that a student may develop mental health issues in the future. These models often leverage longitudinal datasets and behavioral features [9], [10], [19]. Logistic Regression, Decision Trees, and hybrid ensemble models are frequently employed [8], [21]. Risk models are critical because they allow for early intervention strategies aligned with SDG-3 (Good Health and Well-Being).

3) *Treatment Outcome Prediction:*

A smaller but important set of studies attempt to predict the effectiveness of therapeutic interventions [7], [16]. Explainable-AI (XAI) techniques are often used here to ensure transparency and trust in predicting treatment success or therapy adherence [7].

Overall, diagnosis remains the most common objective, but future research trends show growing interest in risk forecasting and treatment personalization. To provide a clearer comparison of how various models perform across different data sources and prediction goals, Table-1 summarizes representative studies along key analytical dimensions.

TABLE I. COMPARATIVE SUMMARY OF PREDICTION OBJECTIVES, DATASET TYPE & AI TECHNIQUES

Ref	Prediction Objective	Dataset Type & AI Technique	Key Performance / Outcome
[5]	Depression Severity	Survey Data & ML (SVM, RF)	Accuracy: 84%, F1-score: 0.79
[15]	Early Detection of Stress	Social media & DL (Transformers)	F1-score: 0.88
[16]	Depression Monitoring	Smartphone Sensor Data & Explainable ML	AUC: 0.91, SHAP for interpretation
[14]	Anxiety & Depression Risk	Wearables + Surveys & Hybrid (CNN + XAI)	Accuracy: 87%, Visual Explanations
[8]	Risk Identification	Clinical + Family Data & ML (Decision Trees)	Precision: 0.76, Recall: 0.81

D. *Challenges Identified:*

Despite ongoing progress, applying AI in mental health analytics still presents key challenges across technical, ethical, and societal dimensions. As shown in the cause-effect diagram in Figure-4, technical issues often relate to data limitations, lack of model generalizability, and interpretability. While models perform well in specific settings, many struggle when applied to broader, diverse populations.

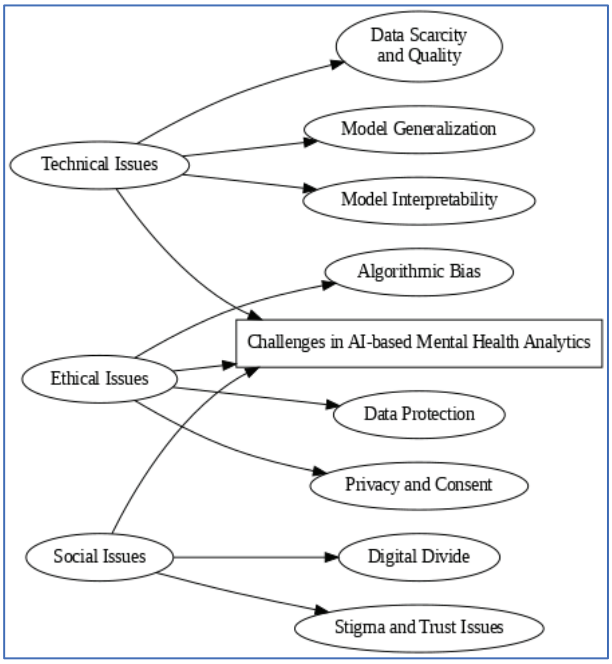

Fig. 4. Fishbone Diagram of Challenges in AI-based Mental Health Analytics

As previously noted in sub-section of Data Sources and Modalities, the limited use of multimodal inputs like voice and image data also restricts emotional context modeling, which could otherwise enhance prediction accuracy. Ethical concerns, especially around data privacy, fairness, and informed consent, remain significant. Although techniques like federated learning are emerging, they are not yet widely adopted. On the social side, access barriers and stigma still limit participation and trust in AI-based tools. These challenges need to be addressed holistically to make AI systems more inclusive, transparent, and scalable supporting long-term goals like SDG-3 and SDG-10.

E. Research Gaps Identified:

The systematic review of 25 articles revealed multiple recurring research gaps that hinder the full potential of AI-driven mental health analytics. These gaps were extracted from the discussion and future work sections of each paper.

1) Lack of Large and Diverse Datasets:

Many studies highlighted that existing datasets are small, imbalanced, and demographically narrow, limiting the generalizability of ML models [2], [5], [8], [11], [19], [22]. Particularly, longitudinal datasets for mental health tracking over time remain scarce.

2) Limited Interpretability and Explainability:

Several DL-based studies reported high accuracy but low interpretability, making it difficult for clinicians to trust and adopt these models [7], [15], [16]. The integration of XAI remains at a nascent stage [11], [16].

3) Over-Reliance on Single Data Modalities:

The majority of studies still rely predominantly on textual data sources, such as surveys or social media posts [5], [8], [15]. The underutilization of multimodal data (voice, sensors, images) restricts the richness of predictive models [2], [6], [11].

4) Lack of Real-World Validation:

Most models were evaluated only in laboratory settings without real-world pilot testing or clinical deployment [9], [18], [23]. Bridging the gap between theoretical performance and real-world usability remains a major challenge.

5) Ethical and Privacy Concerns:

Although many papers acknowledge privacy risks, few studies implement privacy-preserving methods such as differential privacy or federated learning [5], [11], [16], [24].

The identified research gaps highlight critical directions for future research, particularly emphasizing dataset expansion, enhancing model explainability, leveraging multimodal data, ensuring real-world validation, and integrating ethical safeguards in AI systems supporting mental health.

F. SDG Mapping:

The integration of AI into mental health analytics for students supports several SDGs outlined by the United Nations. The 25 reviewed studies predominantly contribute to SDG-3 (Good Health and Well-Being), SDG-4 (Quality Education), and SDG-10 (Reduced Inequalities).

1) Alignment with SDG-3:

Good Health and Well-Being: The majority of studies aimed to enhance mental health diagnosis, monitoring, and early intervention through predictive models [1], [5], [8], [15], [19]. By improving early detection of depression, anxiety, and stress, these research efforts directly contribute to promoting better health and well-being outcomes.

2) Alignment with SDG-4:

Quality Education: Some studies emphasized reducing the academic and psychological burden on students, thereby indirectly improving learning outcomes [4], [8], [10], [18]. Mental health support systems, if implemented effectively, can enhance students' educational achievements and overall academic performance.

3) Alignment with SDG-10:

Reduced Inequalities: A subset of research works targeted privacy-preserving AI methods [5], [11], [16] and context-sensitive models that ensure equitable access to mental health services across diverse demographic groups [19], [23]. These approaches help bridge gaps in mental healthcare accessibility, particularly for marginalized populations.

Thus, AI-driven mental health analytics for students supports a comprehensive sustainable development agenda, enhancing individual well-being, improving educational outcomes, and promoting social equity. This mapping illustrates that research efforts align predominantly with SDG-3 (Good Health and Well-being) while also advancing SDG-4 (Quality Education) and SDG-10 (Reduced Inequalities), contributing to a resilient and inclusive global society.

IV. Evaluation of Current Challenges, Methodologies & Research Gaps

Despite notable advancements in AI-based cybersecurity, several persistent challenges constrain the effectiveness, generalizability, and deployability of current models. This section synthesizes the critical limitations identified across the reviewed literature, with a focus on following three principal issues:

A. Synthesis of Analytical Insights:

Thematic analysis of the reviewed articles highlighted consistent patterns in research focus, technical approaches, and outcome domains. ML techniques dominated the landscape for depression and anxiety diagnosis [5], [8], [15], whereas DL was increasingly employed for unstructured data like voice and social media posts [2], [15], [16].

Hybrid models that integrate ML, DL, and NLP approaches have begun emerging, aiming to improve prediction accuracy and interpretability [11], [14]. Textual data remains the most prevalent modality [2], [5], [8], [15], although growing efforts are visible in leveraging voice, clinical, and sensor data [2], [6], [11]. However, challenges like limited dataset diversity [5], lack of interpretability [7], and inadequate real-world testing [18], [23] persist.

B. Inter-relationships between Methods, Challenges, and Outcomes:

Cross-mapping the reviewed articles shows that, ML techniques (e.g., SVM, RF) often suffer from bias and generalization issues, especially with small datasets [5], [8]. DL models excel in feature extraction but pose explainability challenges, limiting clinical adoption [15], [16]. Risk prediction tasks frequently encounter lack of longitudinal data, impacting prediction validity [9], [10], [19]. Diagnosis models are largely trained on single modality text data, risking oversimplification of mental health conditions [5], [8]. Thus, model type selection, data modality, and target outcome are tightly intertwined with systemic limitations.

C. Emerging Best Practices:

Several best practices are progressively gaining traction in recent studies:

(i) Explainable Models: Incorporating attention mechanisms, SHapley Additive exPlanations (SHAP) values, and interpretable ML pipelines [11], [16].

(ii) Multimodal Data Fusion: Combining textual, acoustic, and physiological signals to enhance robustness [2], [6], [11].

(iii) Privacy-Preserving AI: Introduction of federated learning frameworks for mental health data analysis [11].

(iv) Ethical Compliance: Stressing informed consent, differential privacy, and bias mitigation strategies [5], [11], [16].

These practices directly address technical and ethical bottlenecks and provide blueprints for responsible innovation.

D. Proposed Solutions for Identified Challenges:

Based on the analysis, the following actionable solutions are proposed in Table-2.

TABLE II. PROPOSED SOLUTIONS FOR IDENTIFIED CHALLENGES

Challenge	**Proposed Solutions**
Small Datasets	Synthetic data generation, federated data sharing frameworks
Interpretability Issues	Integration of Explainable-AI methods (SHAP, LIME)
Single Modality Dependence	Multimodal sensor-text fusion
Lack of Real-World Validation	Longitudinal pilot deployments in real educational setups
Privacy and Ethics	Federated learning, differential privacy, ethical audits

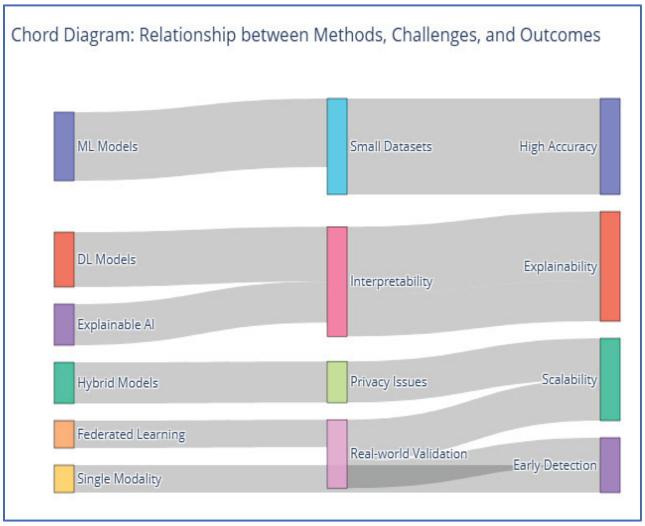

Fig. 5. Chord Diagram: Methods, Challenges & Outcomes

Figure-5, presents a chord diagram illustrating the interconnections between AI methods, challenges, and desired outcomes in mental health analytics. Methods such as ML, DL, and hybrid approaches contribute differently to the emergence of challenges including small dataset availability, interpretability concerns, and privacy risks. These challenges, in turn, act as barriers to achieving key outcomes such as high predictive accuracy, model explainability, and early detection capabilities. The diagram highlights the critical role of overcoming methodological challenges to enable effective, ethical, and sustainable mental health solutions.

Proposing these solutions enhances AI systems' readiness for clinical, educational, and social deployments, reinforcing alignment with SDG-3, 4, and 10. Data diversification and real-world validations remain the most urgent research needs. Federated explainable architectures represent the most promising future direction. This evaluation sets the foundation for the proposed novel conceptual framework described in the next section.

V. PROPOSED NOVEL CONCEPTUAL FRAMEWORK: FEDERATED EXPLAINABLE MENTAL HEALTH ANALYTICS (FEMHA)

The proposed Federated Explainable Mental Health Analytics (FEMHA) framework is designed to directly tackle the key challenges identified during our literature review. FEMHA brings together privacy-preserving federated learning, explainable AI modules, and SDG-aligned outcomes to create a responsible and practical system for predicting mental health risks among students.

A. Key Components:

The FEMHA framework uses a multi-layered structure, where each layer is built to address important needs like privacy, transparency, scalability, and alignment with broader sustainability goals.

1) Multimodal Data Collection Layer:

This is the foundational layer that gathers different types of data critical for strong mental health predictions. It collects following data:

- Text data from surveys, questionnaires, and mental health self-assessments,
- Voice recordings from audio diaries or mobile applications,
- Sensor data from wearables that track metrics like heart rate, sleep, and activity,
- Clinical records from institutional healthcare centers or prior medical histories.

By combining these multiple data sources [2], [6], [8], FEMHA ensures that predictive models are trained on rich, diverse datasets, helping improve accuracy and generalizability across student populations

2) Federated Learning Layer (Privacy-First Training):

To protect sensitive information, FEMHA uses federated learning methods:

- Local model training happens directly on devices or institutional servers, without sending raw data to a central cloud [11], [16].
- Updates from local models are securely aggregated using techniques like Federated Averaging or Federated Optimization.

Additional methods such as differential privacy and secure multiparty computation can also be applied to make the system even safer. This layer ensures that collaboration happens without compromising data ownership or student privacy.

3) Explainable AI Decision Layer:

Since trust is essential in mental health applications, FEMHA includes a strong explainability component:

- Techniques like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and attention-based networks are used to make model predictions more understandable [7], [15].
- Clinicians, educators, and even students themselves can see the factors influencing a prediction.

If an unusual prediction is made, it can be flagged and explained, supporting fairness and ethical use of AI. This helps ensure that FEMHA's outputs are not just black-box results but meaningful, human-centered insights.

4) SDG-Driven Outcome Optimization Layer:

Finally, FEMHA connects its predictive work directly to global development goals:

- SDG-3 (Good Health and Well-being): Early detection and intervention for better student mental health.
- SDG-4 (Quality Education): Helping students stay engaged and succeed academically.
- SDG-10 (Reduced Inequalities): Making sure mental health support reaches students across different backgrounds fairly [18], [19].

FEMHA measures success not just by prediction accuracy, but also by fairness, energy efficiency, and its contribution to sustainable development.

B. Working Mechanism:

The operational flow of FEMHA follows a clear, privacy-conscious, and sustainability-focused process. The step-by-step working process is given below:

Step 1: Multimodal Data Capture: Students interact with apps, wearable devices, and institutional systems to generate different types of data such as Survey responses and textual assessments, Voice recordings through mobile apps, Physiological data collected through wearables, Clinical records from health centers. All of this data stays securely stored on local servers or personal devices, respecting privacy laws and ethical standards.

Step 2: Local Model Training with Federated Learning: Instead of sending sensitive data to a central server, each device or institution trains its own local AI model. Only model updates (like weight changes) are shared, keeping student information secure while still contributing to better learning models across the network.

Step 3: Secure Model Aggregation and Global Updates: The central aggregator node collects the updates from all local models and securely merges them. Methods like Federated Averaging allow the system to build a strong, general model without ever touching raw student data. This step also ensures that the system stays robust even when different institutions have different types of data.

Step 4: Explainable Decision-Making: Before making final mental health predictions, FEMHA applies explainability tools. Methods like SHAP, feature contribution ranking, or attention heatmaps show why a particular prediction was made. This step builds transparency, builds user trust, and helps clinicians or educators make better decisions based on the AI's suggestions.

Step 5: Sustainable Outcome Monitoring and SDG Alignment: Finally, FEMHA monitors how well the system is doing by checking against three major goals – (i) - Health improvements for students (aligned with SDG-3), (ii) - Better educational outcomes through improved academic support (aligned with SDG-4), (iii) - Fairness and accessibility across different student groups (aligned with SDG-10).

Feedback loops are built into the system, so it can keep adapting and improving its performance over time in real-world settings.

The flowchart in Figure-6, illustrates the sequential operation of the FEMHA framework, beginning with multimodal data capture and proceeding through local training, secure aggregation, explainable decision generation, and SDG-aligned outcome monitoring. Each stage maintains a privacy-first, sustainable, and ethical design.

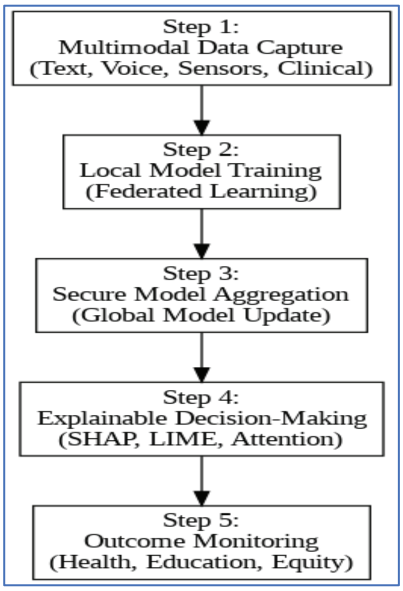

Fig. 6. Flowchart of FEMHA Framework Operational Workflow

C. Justification and Addressing Literature Gaps:

The FEMHA Framework addresses the following key literature gaps, listed in Table-3.

TABLE III. FEMHA CONTRIBUTION TO ADDRESS LITERATURE GAPS

Identified Gap	FEMHA Contribution
Privacy Risks	Federated Learning avoids central data storage [11], [16].
Lack of Explainability	Embedded XAI modules provide model transparency [7], [15].
Single Modality Dependence	Multimodal fusion enriches feature representation [2], [6].
Lack of Real-World Validation	Designed for deployment in decentralized real-world environments (universities, colleges).
Equity Challenges	SDG alignment ensures inclusion of marginalized groups [18], [23].

Thus, FEMHA represents a holistic, sustainable, and ethically responsible mental health AI system.

D. Sustainability and Future-Readiness:

The FEMHA Framework offers following benefits related to it's sustainability and future-readiness.

1) Energy-Efficient Training:

Computational overhead is decreased via lightweight federated systems.

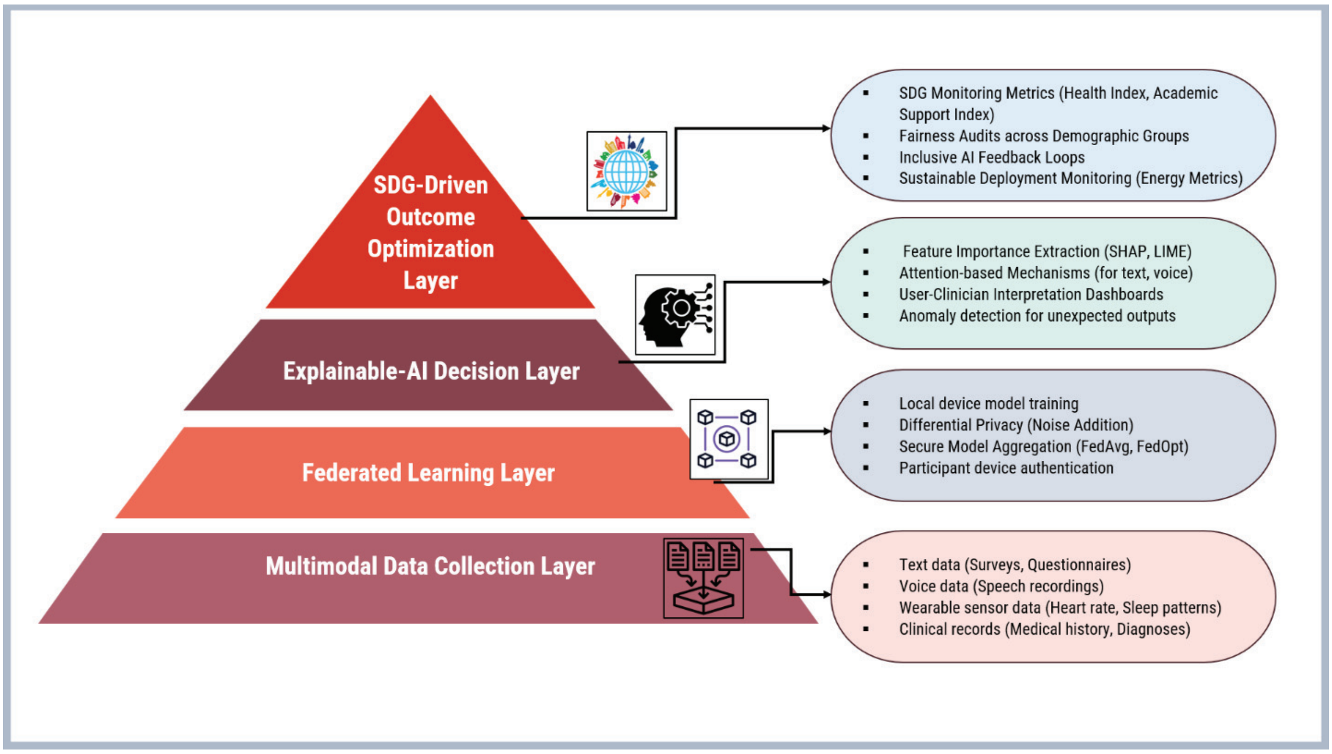
SDG-Driven Outcome Optimization Layer
Explainable-AI Decision Layer
Federated Learning Layer
Multimodal Data Collection Layer
SDG Monitoring Metrics (Health Index, Academic Support Index)
Fairness Audits across Demographic Groups
Inclusive AI Feedback Loops
Sustainable Deployment Monitoring (Energy Metrics)
Feature Importance Extraction (SHAP, LIME)
Attention-based Mechanisms (for text, voice)
User-Clinician Interpretation Dashboards
Anomaly detection for unexpected outputs
Local device model training
Differential Privacy (Noise Addition)
Secure Model Aggregation (FedAvg, FedOpt)
Participant device authentication
Text data (Surveys, Questionnaires)
Voice data (Speech recordings)
Wearable sensor data (Heart rate, Sleep patterns)
Clinical records (Medical history, Diagnoses)

Fig. 7. Layered Pyramid Diagram - FEMHA Framework & Components

2) Support for Underrepresented Groups:

SDG-10 alignment is ensured by demographic stratification, inclusion policies, and fairness audits.

3) Adaptability:

Able to change to meet new AI standards (explainable transformers, technologies that improve privacy).

Figure-7, illustrates the layered pyramid view FEMHA framework's hierarchical architecture, which begins with data collection at the base and progresses through explainable decision-making, privacy-preserving learning, and SDG-driven optimized results.

The FEMHA Framework offers a futuristic blueprint for AI-driven mental health analytics, ensuring privacy, explainability, equity, and sustainability, setting a new benchmark for research at the intersection of AI, healthcare, and education.

VI. Conclusion and Future Directions

A. Summary of Key Findings:

This study reviewed twenty-five peer-reviewed articles to understand how AI is being used to support mental health prediction, especially for students facing depression, stress, and anxiety. The findings show that AI-based techniques such as ML, DL, hybrid models, and XAI have helped improve prediction accuracy, identify key risk factors, and support early interventions.

However, several common challenges were observed across many studies:

- Small datasets, which often lead to overfitting [5], [8]
- Limited explainability in deep models, making it harder for clinicians to trust them [7], [15]
- Privacy risks arise from the central storage of sensitive data [11], [16]
- Lack of real-world validation hinders large-scale adoption [18], [23]

B. Future Scope of Research:

Based on the gaps and promising practices identified in this review, several directions are recommended for future work:

1) Real-World Pilots:

AI models should be tested in real educational or healthcare environments.

2) Work Across Disciplines:

Future research should integrate mental health professionals, ethicists, and educators.

3) Adopting Diverse Data:

Adding behavioral and sensor-based data through wearable devices to traditional surveys will enhance the model reliability and provide enhanced predictions.

4) Design Transparent and Fair AI:

It is critical to design models that are user-friendly, treat all groups fairly, and clearly explain their predictions. This will increase trust and promote wider use.

5) Focus on Sustainability:

As AI tools are deployed in real settings, especially on mobile or edge devices, models should be lightweight and energy-efficient to ensure long-term sustainability.

The flowchart in Figure -8, outlines the strategic roadmap for advancing AI-driven mental health research, emphasizing real-world deployment, ethical compliance, interdisciplinary collaboration, and SDG alignment.

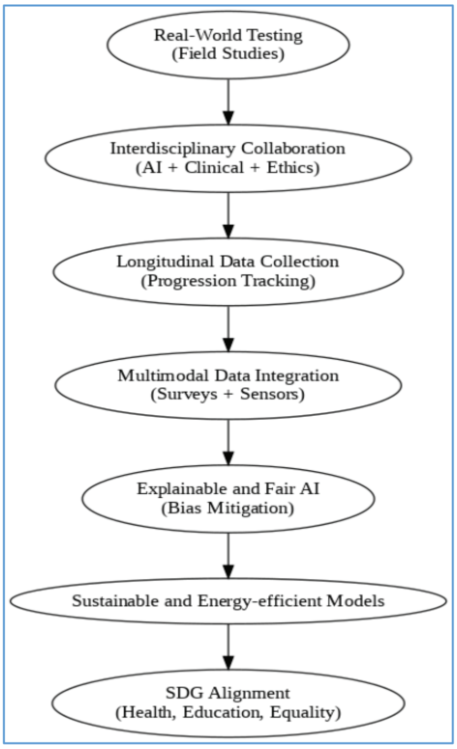
Real-World Testing
(Field Studies)
Interdisciplinary Collaboration
(AI + Clinical + Ethics)
Longitudinal Data Collection
(Progression Tracking)
Multimodal Data Integration
(Surveys + Sensors)
Explainable and Fair AI
(Bias Mitigation)
Sustainable and Energy-efficient Models
SDG Alignment
(Health, Education, Equality)

Fig. 8. Roadmap for Future Research in AI-Driven Mental Health Analytics

In conclusion, responsible AI-driven mental health analytics promises to redefine preventive mental health care by merging technical excellence with ethical responsibility, ensuring a safer, healthier, and more inclusive future for students worldwide.

References

[1] A. Dawood, S. Turner, and P. Perepa, “Affective Computational Model to Extract Natural Affective States of Students With Asperger Syndrome (AS) in Computer-Based Learning Environment,” IEEE Access, vol. 6, pp. 67026–67034, 2018, doi: https://doi.org/10.1109/access.2018.2879619.

[2] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, “Mental health monitoring with multimodal sensing and machine learning: A survey,” Pervasive and Mobile Computing, vol. 51, pp. 1–26, Dec. 2018, doi: https://doi.org/10.1016/j.pmcj.2018.09.003.

[3] S. Festag and C. Spreckelsen, "Privacy-Preserving Deep Learning for the Detection of Protected Health Information in Real-World Data: Comparative Evaluation," JMIR Formative Research, vol. 4, no. 5, p. e14064, May 2020, doi: https://doi.org/10.2196/14064

[4] C. Wang, H. Zhao, and H. Zhang, "Chinese College Students Have Higher Anxiety in New Semester of Online Learning During COVID-19: A Machine Learning Approach," Frontiers in Psychology, vol. 11, p. 587413, 2020, doi: https://doi.org/10.3389/fpsyg.2020.587413

[5] A. Priya, S. Garg, and N. P. Tigga, “Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms,” Procedia Computer Science, vol. 167, pp. 1258–1267, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.442.

[6] T. Richter, B. Fishbain, A. Markus, G. Richter-Levin, and H. Okon-Singer, “Using machine learning-based analysis for behavioral differentiation between anxiety and depression,” Scientific Reports, vol. 10, no. 1, p. 16381, Oct. 2020, doi: https://doi.org/10.1038/s41598-020-72289-9.

[7] S. Hornstein, V. Forman-Hoffman, A. Nazander, K. Ranta, and K. Hilbert, “Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach,” DIGITAL HEALTH, vol. 7, p. 205520762110606, Jan. 2021, doi: https://doi.org/10.1177/20552076211060659.

[8] M. Gil, S. S. Kim, and E. J. Min, "Machine Learning Models for Predicting Risk of Depression in Korean College Students: Identifying Family and Individual Factors," Frontiers in Public Health, vol. 10, p. 1023010, Nov. 2022, doi:https://doi.org/10.3389/fpubh.2022.1023010

[9] R. Qasrawi, S. Vicuna Polo, D. Abu Al-Halawah, S. Hallaq, and Z. Abdeen, “Schoolchildren’ Depression and Anxiety Risk Factors Assessment and Prediction: Machine Learning Techniques Performance Analysis (Preprint),” JMIR Formative Research, Aug. 2021, doi:https://doi.org/10.2196/32736.

[10] L. Liang, Y. Zheng, Q. Ge, and F. Zhang, “Exploration and Strategy Analysis of Mental Health Education for Students in Sports Majors in the Era of Artificial Intelligence,” Frontiers in Psychology, vol. 12, Mar. 2022, doi: https://doi.org/10.3389/fpsyg.2021.762725.

[11] M. S. Ahmed and N. Ahmed, “A Minimal and Faster System to Identify Depression Through Smartphone: An Explainable Machine Learning-Based Approach (Preprint),” JMIR Formative Research, Dec. 2022, doi: https://doi.org/10.2196/28848.

[12] A. Baba and K. Bunji, “Prediction of Mental Health Problem Using Annual Student Health Survey: A Machine Learning Approach (Preprint),” JMIR Mental Health, Sep. 2022, doi: https://doi.org/10.2196/42420.

[13] H. Lyu, “Application of machine learning on depression prediction and analysis,” Applied and computational engineering, vol. 5, no. 1, pp. 712–719, Jun. 2023, doi: https://doi.org/10.54254/2755-2721/5/20230681.

[14] Hanif Abdul Rahman et al., “Machine Learning-Based Prediction of Mental Well-Being Using Health Behavior Data from University Students,” Bioengineering, vol. 10, no. 5, pp. 575–575, May 2023, doi: https://doi.org/10.3390/bioengineering10050575.

[15] Biodoumoye George Bokolo and Q. Liu, “Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques,” Electronics, vol. 12, no. 21, pp. 4396–4396, Oct. 2023, doi: https://doi.org/10.3390/electronics12214396.

[16] D. Liu, Z. Chen, W. J. Marrero, N. C. Jacobson, and T. Thesen, “Explainable machine learning-based prediction of depression severity in medical students,” Dec. 2023, doi: https://doi.org/10.1101/2023.12.14.23299975.

[17] F. Norouzi and B. L. M. Santos Machado, "Predicting Mental Health Outcomes: A Machine Learning Approach to Depression, Anxiety, and Stress," International Journal of Applied Data Science in Engineering and Health, vol. 1, no. 2, Oct. 2024

[18] A. Amin, Liza, N. Shah, N. E. Hasan, M. Musa, and J. Chakra, “Predicting and Monitoring Anxiety and Depression: Advanced Machine Learning Techniques for Mental Health Analysis,” British Journal of Nursing Studies, vol. 4, no. 2, pp. 66–75, Oct. 2024, doi: https://doi.org/10.32996/bjns.2024.4.2.8.

[19] Y. Zhai et al., “Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students,” Journal of Counseling & Development, Oct. 2024, doi: https://doi.org/10.1002/jcad.12543.

[20] A. Yoo et al., “Prediction of adolescent depression from prenatal and childhood data from ALSPAC using machine learning,” Scientific Reports, vol. 14, no. 1, Oct. 2024, doi: https://doi.org/10.1038/s41598-024-72158-9.

[21] D. K. Saha, T. Hossain, M. Safran, Sultan Alfarhood, M. F. Mridha, and D. Che, “Ensemble of hybrid model based technique for early detecting of depression based on SVM and neural networks,” Scientific Reports, vol. 14, no. 1, Oct. 2024, doi: https://doi.org/10.1038/s41598-024-77193-0.

[22] Muhammad Azizur Rahman and T. Kohli, “Mental health analysis of international students using machine learning techniques,” PloS one, vol. 19, no. 6, pp. e0304132–e0304132, Jun. 2024, doi: https://doi.org/10.1371/journal.pone.0304132.

[23] M. T. Mardini, G. E. Khalil, C. Bai, Aparna Menon DivaKaran, and J. M. Ray, “Identifying Adolescent Depression and Anxiety through Real-world Data and Social Determinants of Health (Preprint),” JMIR Mental Health, vol. 12, pp. e66665–e66665, Dec. 2024, doi: https://doi.org/10.2196/66665.

[24] Y. Fu, F. Ren, and J. Lin, “Apriori algorithm based prediction of students’ mental health risks in the context of artificial intelligence,” Frontiers in Public Health, vol. 13, Feb. 2025, doi: https://doi.org/10.3389/fpubh.2025.1533934.

[25] L. Yang et al., “Application of machine learning in depression risk prediction for connective tissue diseases,” Scientific Reports, vol. 15, no. 1, Jan. 2025, https://doi.org/10.1038/s41598-025-85890-7.