# A Privacy-Aware Federated Hybrid Model for Multimodal Mental Health Analysis

Yusra, Riaz UlAmin
Smart Systems Research Lab, University of Okara, Pakistan

***Abstract*—As mental health disorders such as stress, anxiety, depression, and post-traumatic stress disorder (PTSD) affect a substantial part of the world population, current diagnostic methodologies are still centralized, subjective, and sensitive to privacy concerns. To mitigate these limitations, this study presents a new framework for multimodal mental health classification within a privacy-preserving federated learning framework learning using electroencephalography (EEG), electrocardiography (ECG) and galvanic skin response (GSR) signals. Furthermore, we propose a hybrid deep learning architecture, which combines CNN-LSTM-Transformer blocks to effectively learn spatial, temporal, and long-range dependencies within physiological signals.After preprocessing the cleaned data through artifact removal, band-pass filtering, normalization and multimodal feature fusion signal quality is improved. The proposed model is trained in a federated setting with multiple clients for decentralized training without sharing raw data allowing it to preserve privacy and communication efficiency supporting non-IID data extensions. We evaluate on two datasets, SAM40 (stress detection) and DAPS (anxiety, depression, and PTSD classification). The proposed framework achieved 97% accuracy on SAM40 and more than 96% accuracy on DAPS. Comparative assessments with recent federated and centralized methods validate its strength in multimodal fusion and robust feature exploitation. These results demonstrate the possibility of a general framework for designing privacy-preserving and efficient mental health monitoring systems that can support both Clinical and Wearable-device applications.**

***Keywords*—Mental health detection; hybrid CNN-LSTM-transformer; federated learning; multimodal physiological signals**

## I. Introduction

Mental health illnesses, like depression, stress, and anxiety, affect one billion people each year and create significant social, economic, and healthcare burdens. Depression, stress, and anxiety disorders are some of the top causes of disability, and loss of healthy and productive years of life [1]. Depression is a major cause of suicide and loss of work productivity and costs the world economy 1 trillion dollars a year [2].

Mental disorders must be identified and assessed to provide intervention. Current methods, such as interviews interviews and self-reported questionnaires are subjective, time-consuming, and impractical. Measuring physiological states can be used to indicate mental health because they are objective. Physiological signals continuously monitored and measured. EEG records brain activity, ECG measures heart activity due to stress or anxiety, and GSR measures the arousal from the autonomic nervous system [3]. While machine learning techniques can efficiently analyze given problems, centralized methods require collecting personal and sensitive data, raising serious privacy concerns. Another important point is that most studies focus on analyzing a single type of signal, and ignores other potentially useful signals. Real-world data that can be collected from hospitals, clinics, and wearable devices are also non-independent and identically distributed (or non-IID), which makes it difficult for models to generalize effectively [4].

To mitigate those problems, advanced preprocessing and feature extraction is done before federated training. EEG signals are filtered to a desired band, artifacts are removed, and other signals are divided into segments. For processed ECG signals, we determine R-peaks and extract important features such as heart rate and heart rate variability (HRV) from our filtered signals. GSR signals are smoothed and normalized, and we identify arousal peaks. These methods minimize added noise in data, make the characteristics and inputs we are analyzing uniform, and improve feature extraction through cross-feature analysis in time and space across different signals. [5].

Federated Learning (FL) allows for training models in a distributed manner across devices in a privacy-preserving manner, which allows which allows for better compliance with privacy regulations. Using multiple modalities in a federated system improves detection performance by enabling the capture of spatial, temporal, and cross-modal dependencies. We present a novel CNN-LSTM-Transformer architecture where the convolutional layers capture the spatial patterns, the LSTM layers capture the temporal patterns, and the Transformer layers capture long-range and cross-modal dependencies. The framework has capacity to further integrate privacy-preserving techniques such as Differential Privacy and Secure Aggregation to help keep distributed training secure and ensure that model updates are safely combined in the federated learning setting [6].

As illustrated in Fig. 1, the AI system supports anxiety disorder management using multiple modalities.

The main contributions of this work are:

- We explore a multimodal federated learning framework integrating EEG, ECG, and GSR for enhanced and robust mental health detection.
- We developed a CNN-LSTM-Transformer model to illustrate hierarchical feature extraction from preprocessed physiological signals.
- We remove noise and artifacts from EEG, ECG, and GSR signals to improve signal integrity and feature reliability.
- We comprehensively evaluate the proposed approach in terms of accuracy, F1-score, convergence rate, communication efficiency, and privacy preservation.

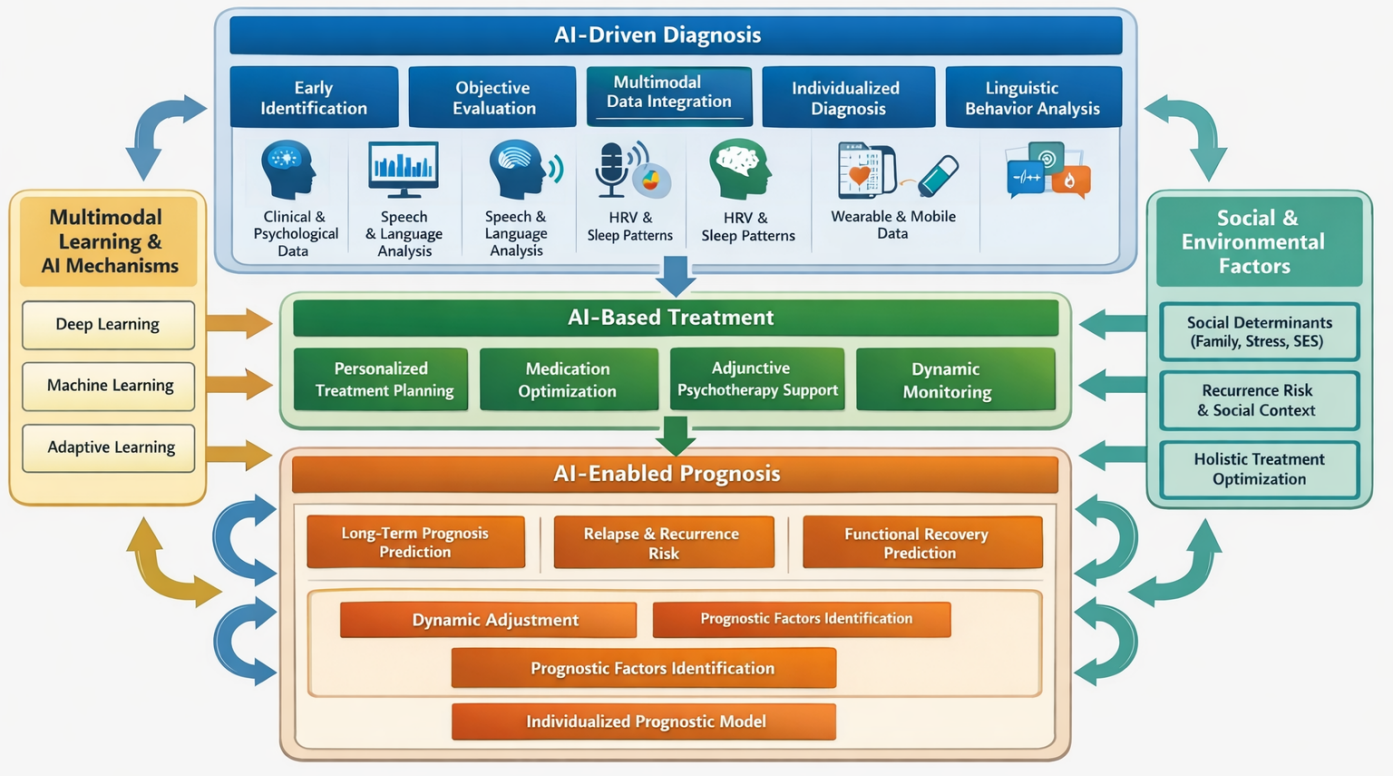


Fig. 1. AI-Driven conceptual framework for anxiety disorder integrating diagnosis, treatment, and prognosis with multimodal data, machine learning, and social-environmental factors.

## II. Literature Review

Recently, federated learning (FL) has emerged to enhance privacy-preserving mental health and emotion recognition systems by enabling decentralized training without sharing raw physiological data. Although there have been these advances, the majority of the current studies still have drawbacks in terms of multimodal fusion, robustness against non-IID data distribution and scalability for real clinical applications. In their proposed privacy-preserving federated framework for depression relapse prediction from EEG signals and clinical symptom scores, Yasin et al. [7] introduced a method that leverages machine learning to predict depression relapse by analyzing electroencephalogram (EEG) signals and symptom scores. The model uses explainable AI methods like Layer-wise Relevance Propagation (LRP) and Shapley values, and it has an accuracy of 92%. However, the method does not consider multimodal physiological integration, and is only applied to single-modality EEG data.

Dubey et al. [8] proposed a multimodal federated learning system with CNN and LSTM networks under various privacy constraints. The method remains sensitive to data heterogeneity and is not very adaptive to various client distributions as it achieves only around 92% accuracy. Simic et al. [9] presented a federated multimodal emotion recognition system based on CNN models for audio and visual feature extraction. The model uses decision-level fusion, which restricts the ability of deep cross-modal feature interaction resulting in limited affective state modeling performance, although FedAvg can enhance the generalization of the model.

Gupta et al. [10] proposed an EEG-audio-based federated multimodal system for the early detection of the Major Depressive Disorder using Bi-LSTM based fusion. Although the model performed well with 91.9% accuracy, the extent to which the model has been validated on real-world heterogeneous datasets is limited, which results in a lack of clinical generalization.

Almadhor et al. [11] studied stress detection in the context of wearable electrodermal activity (EDA) signals in a federated learning framework. The system provides 86.82% accuracy but under device variations in the devices and limited diversity in the datasets, the accuracy of the system drops.

Kumar et al. [12] proposed a federated transfer learning approach (FTL-Emo) for EEG-based emotion recognition. The model obtains an F1-score of 92-94%, but does not generalize well across datasets due to distribution shifts between the training and testing domain. Federated reinforcement learning combined with machine unlearning was used in healthcare monitoring and anomaly detection by Shaik et al. [13]. The framework offers flexibility and privacy, but it has high computational complexity, which restricts its use in real-time applications.

Farooq et al. [14] proposed a federated learning approach involving local classifiers with a meta-classifier for the detection of autism spectrum disorders. When applied to children, the method achieved 98% accuracy, but decreased to 81% for adults, thereby showing poor generalisation across age groups.

More recently, Pradeep et al. [15] investigated cross-modal

TABLE I. COMPREHENSIVE LITERATURE REVIEW OF FEDERATED LEARNING AND HYBRID AI APPROACHES IN MENTAL HEALTH AND EMOTION RECOGNITION

| Reference | Objective | Methodology | Input Data | Hybrid | AI/ML | Limitations | Accuracy (%) | Bench-marked |
|---|---|---|---|---|---|---|---|---|
| [7] | Early prediction of depression relapse with privacy preservation | Federated learning with EEG features and clinical scores; explainable AI | EEG + symptom scores | ✓ | ✓ | Computational demand, EEG quality dependent | 92 | ✓ |
| [8] | Privacy-preserving mental health prediction | Federated learning with CNN + LSTM, multimodal data integration, privacy | EEG, ECG, GSR, behavioral signals | ✓ | ✓ | Data heterogeneity, privacy trade-offs | 92 | ✓ |
| [9] | Privacy-preserving multimodal emotion recognition on edge devices | CNN-based audio + video model with federated learning (FedAvg) | Facial video frames + Audio spectrograms | ✓ | ✓ | Sensor variability, model interpretability | 92.86–94.05 | ✓ |
| [10] | Early MDD detection with privacy | Federated multimodal deep learning (EEG + audio, Bi-LSTM) | EEG + audio signals | ✓ | ✓ | Limited real-world validation | 91.9 | ✓ |
| [11] | Privacy-preserving stress detection | Federated DNN (with SMOTE and normalization) | Wearable electrodermal activity (WESAD dataset) | ✓ | ✓ | Limited sample size, device variability | 86.82 | ✓ |
| [12] | Privacy-preserved automatic emotion recognition | Federated transfer learning (CNN + FL) | EEG biomarkers (DEAP + K-EmoCon) | ✓ | ✓ | Limited cross-dataset generalization | 92–94 | ✓ |
| [13] | Healthcare monitoring using FRL and machine unlearning | Federated reinforcement learning (FedStack, FRAMU) | Multimodal sensor | ✓ | ✓ | Computational complexity, data heterogeneity | - | ✓ |
| [14] | Early ASD detection with privacy | Federated learning with local LR + SVM and meta-classifier | Four ASD feature datasets | ✓ | ✓ | Limited adult dataset generalization | 98 (children), 81 (adults) | ✓ |
| [15] | Cross-modal brain–heart interaction modeling using EEG and ECG feature mapping | EEG–ECG cross-modal feature learning with regression-based mapping | EEG + ECG signals | ✓ | ✓ | No federated learning, no privacy-preserving framework, limited scalability | 93.4 | ✗ |
| **Proposed Method** | Hybrid CNN-LSTM-Transformer with noise removal for anxiety detection | CNN, LSTM, Transformer, artifact removal, noise filtering | EEG ,ECG,GSR signals | ✓ | ✓ | Computationally intensive, EEG variability | 97.37 | ✓ |

brain–heart interaction modeling using EEG and ECG signals. The study was found to be effective for physiological feature mapping across modalities, but lacks of federated learning or privacy-preserving distributed training, making it less applicable in real-world decentralized healthcare settings.

The literature summarized in Table I shows that recent studies have explored the use of federated learning techniques for mental health and emotion recognition. Although current approaches help privacy of data, face challenges such as limited generalization, weak multimodal fusion, data sensitivity, and lack of clinical validation in real-world applications.

The proposed method, however, combines the three signals — EEG, ECG and GSR through a hybrid CNN–LSTM–Transformer network with improved feature fusion and noise reduction. This enables the model to capture spatial, temporal and long-range dependencies and maintain robustness in different federated environments, that increases the performance and scalability of real-world mental health monitoring systems.

## III. RESEARCH METHODOLOGY

This research develops a novel, multi-modal, federated learning based, mental health detection framework that prioritizes and balances privacy and scalability. The system fuses EEG, ECG, and GSR data and utilizes Convolutional Neural Networks, Long Short-Term Memory networks and Transformers, allowing to capture the spatio-temporal and attention driven features. This staircase approach facilitates the overall characterization and modeling of stress, anxiety, and depression related multi-dimensional physiological responses.

### A. Data Collection and Understanding

In this research, the DAPS and SAM40 datasets, which have been incorporated into clinical and wearable health monitoring systems, involve several datasets sampling a wide range of mental states and multiple-spectral modalities.

The SAM40 dataset contains EEG wave recordings of 40 participants who were volunteers in a convenience sample of a greater research project (14 females, 26 males; average age 21.5 years) who were treated to a series of cognitive task activations designed to produce and control stress states. Tasks included were Stroop activities, mental arithmetic, mirror image processing and formal relaxation. Active data collection utilized a 32-channel Emotiv Epoc Flex EEG headset which adhered to the international 10–20 system spatial configuration, where Fp1, Fp2, F3, F4, C3, C4, O1, O2 and a few other electrodes were recorded. Signals were collected and within the resultant 2.75 minutes of quiet recording, were arranged in 25 second, non-overlapping intervals. Each resultant interval epoch data was represented as matrix [16], [17].

$$X^{(t)} = \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \\ \vdots \\ x_C^{(t)} \end{bmatrix} \in \mathbb{R}^{C \times T}, \quad C = 32,\ T = 25 \times 128 \qquad (1)$$

where, $x_c^{(t)}$ corresponds to the time-series signal from channel $c$ for epoch $t$, with $C$ as the total number of EEG channels and $T$ as the number of samples per epoch. Our collection has both continuous raw and artifact-corrected EEG signals. This feature makes SAM40 a flexible dataset as to the study of brain activities associated with varying stress levels and the use of machine learning tools.

DAPS dataset (custom or proprietary) comprises EEG and GSR signals of GSR and EEG collected from participants having depression, anxiety, or PTSD, as well as healthy controls. EEG is collected with 16–128 channels deploying a high-density clinical or wearable EEG system, GSR is obtained from finger/palm electrodes, and ECG is captured optionally via chest leads or wearable sensors. Data collection occurs during resting-state (eyes open/closed) and emotion or task

induced conditions where clinical scales (e.g. BDI, HAM-A) and interactivity and valence ratings are used to derive labels. Each of the modalities is defined as follows:

$$\mathbf{X}_{EEG}^{(t)} \in \mathbb{R}^{C_{EEG} \times T}, \quad \mathbf{X}_{GSR}^{(t)} \in \mathbb{R}^{C_{GSR} \times T}, \quad \mathbf{X}_{ECG}^{(t)} \in \mathbb{R}^{C_{ECG} \times T} \tag{2}$$

Data are distributed in a heterogeneous, non-IID manner across clients:

$$D = \bigcup_{k=1}^{K} D_k, \quad D_i \cap D_j = \emptyset \text{ for } i \neq j \tag{3}$$

This equation shows that the overall dataset $D$ is the combination of client datasets $D_k$ and that each client's data is disjoint, meaning no clients get to see the same samples. This arrangement mirrors practical federated learning situations with non-IID data distributions.

This organization of datasets assists in providing a strong basis for multimodal mental health detection and allows for the development of privacy-preserving and scalable models.

### *B. Signal Preprocessing and Feature Engineering*

In order to thoroughly assess various mental health conditions based on several physiological signals, a careful range of steps involving a preprocessing pipeline predicated on a theory and the crafting of various features must be devised and executed. Although EEG, ECG, and GSR signals contain substantial amounts of noise and artifacts, they may contain a lot of useful information, especially with regard to inter-subject variance. When performed properly, feature extraction and the noise and artifacts of a signal, termed preprocessing, should be done to obtain a signal that accurately represents the underlying physiological and cognitive states, and this will enable precise modeling.

EEG records electrical activity as generated by postsynaptic potentials due to the synchronous firing of neurons. There are different cognitive and emotional states reflected in varying ranges of delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz) waves. However, it's worth noting that EEG is particularly vulnerable to artifacts due to eye and muscle movement, as well as body motion [18]. This is why band-pass filtering is employed in order to retain meaningful activity within the oscillations. The electrical signals generated by the heart (ECGs) are also filtered in order to remove baseline wander and motion artifacts to more accurately quantify the variability in heart rate (HRV), a well-established marker of stress and anxiety, and of autonomic nervous system activity [19]. GSR signals, which capture changes in skin conductance that are due to nervous system activity, are filtered to extract the stress, emotional reactivity, and arousal-response-associated components. The mathematically conveyed expression associated with the artifact-free signal is:

$$\tilde{x}_c(t) = \mathcal{F}_{\text{filter}}(x_c(t)) \tag{4}$$

where, $\tilde{x}_c(t)$ is the preprocessed signal at channel $c$, $x_c(t)$ is the raw signal, and $\mathcal{F}_{\text{filter}}$ represents the corresponding artifact removal filter.

Physiological data often have samples missing due to sensor malfunctions or motion artifacts. Interpolation is employed to keep the time series continuous, so as to not introduce bias into the statistical and spectral data analyses.

$$x_c(t) = \frac{x_c(t-1) + x_c(t+1)}{2}, \quad \text{if } x_c(t) \text{ is missing} \tag{5}$$

Standardization and normalization are critical to mitigate inter-subject and inter-device variability [20]. Signals are standardized to zero mean and unit variance:

$$\begin{aligned} \tilde{x}_c(t) &= \frac{x_c(t) - \mu_c}{\sigma_c}, \\ \mu_c &= \frac{1}{T}\sum_{t=1}^{T} x_c(t), \\ \sigma_c &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_c(t) - \mu_c)^2} \end{aligned} \tag{6}$$

Here, $\mu_c$ and $\sigma_c$ are the mean and standard deviation for the signal of each channel $c$. Lastly, for additional variability mitigation, normalization can rescale signals to a range of [0,1]:

$$x_c^{\text{norm}}(t) = \frac{x_c(t) - x_c^{\min}}{x_c^{\max} - x_c^{\min}} \tag{7}$$

Feature extraction is based on the realization that psychological conditions are expressed as patterned changes on multiple time scales, spectral compositions, and physiological changes of various durations. Characteristics of the time domain, such as mean, variance, standard deviation, as well as HRV, are required to assess the overall dynamics of a signal:

$$\text{Mean: } \mu = \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \text{Variance: } \sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu)^2 \tag{8}$$

The Fast Fourier Transform (FFT) is utilized to derive frequency-domain characteristics through quantifying power in specific frequency bands related to stress, attention, or relaxation as follows:

$$X(f) = \sum_{t=0}^{T-1} x_t e^{-j2\pi ft/T} \tag{9}$$

Features in the time-frequency domain corresponded to the non-stationary and transient characteristics [21], which are relevant to the modeling of stress and emotions, and are obtained through the use of Discrete Wavelet Transform (DWT) as:

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int x(t)\psi\left(\frac{t-b}{a}\right) dt \tag{10}$$

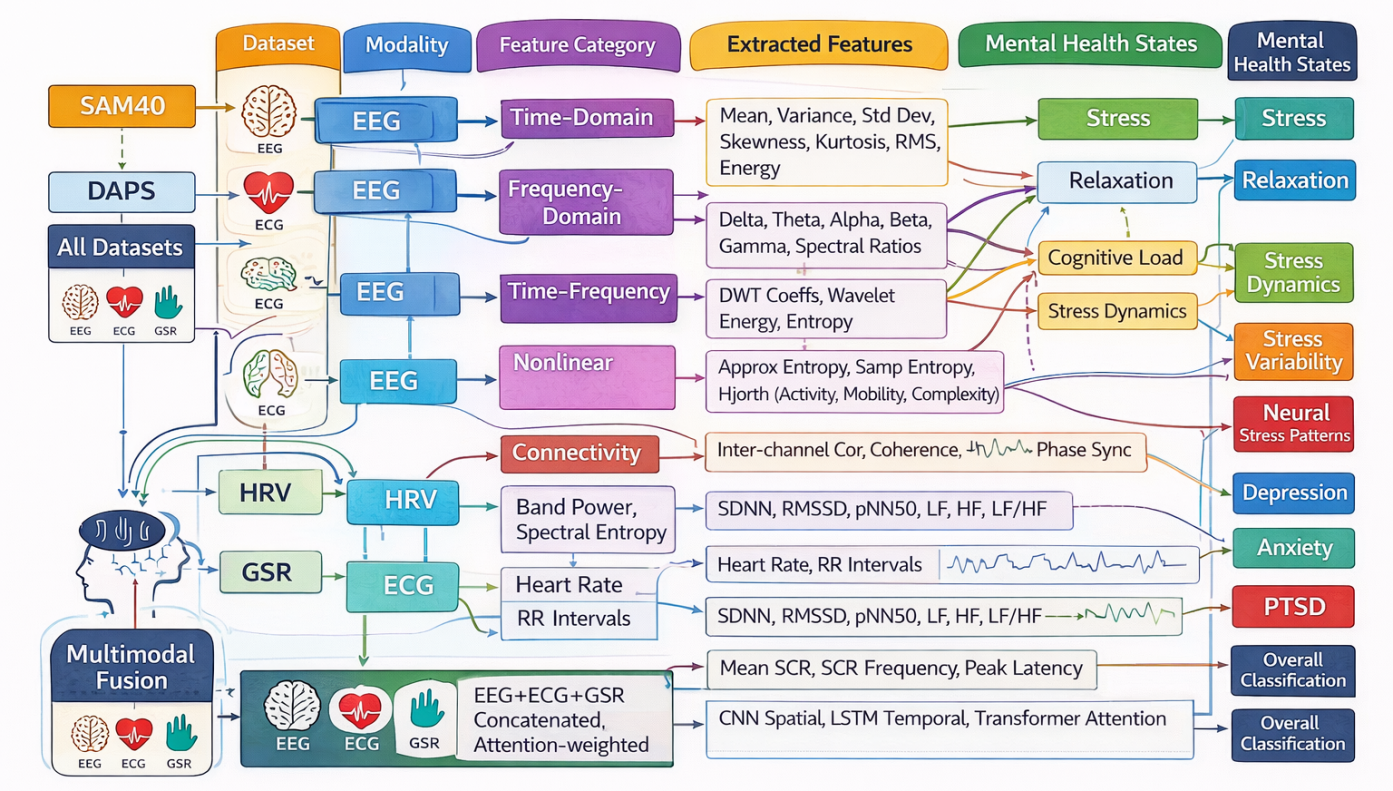


Fig. 2. Correlation of multimodal physiological features with mental health states.

where, $a, b$ are scale and translation respectively and $\psi(t)$ is the mother wavelet.

To take advantage of the complementary information across modalities, the EEG, ECG, and GSR features are concatenated into a single, unified representation:

$$\mathbf{F} = [\mathbf{F}_{EEG}\ \mathbf{F}_{ECG}\ \mathbf{F}_{GSR}] \tag{11}$$

An attention-based weighting mechanism is utilized to emphasize information that is relevant while diminishing information that is redundant or of less relevance:

$$\mathbf{F}_{\text{att}} = \mathbf{F} \odot \alpha, \quad \sum_i \alpha_i = 1 \tag{12}$$

where, $\alpha$ defines learned attention weights and where $\odot$ represents a Hadamard product. This makes sure that the notable physiological patterns contributing to the stress, anxiety, and/ or depression are accentuated in the final feature representation.

There is a feature set providing a rich, complementary, and physiologically meaningful representation due to the combination of artifact-free, uniform, and standardized signals with temporal, spectral, and time-frequency features, which are of distinguished discriminative value, along with the modalities being fused with an attention mechanism. This also strengthens the capabilities of hybrid CNN-LSTM-Transformer models in identifying minute changes in mental health conditions. This makes the technique highly recommended for privacy-preserving federated learning systems as it is as applicable as possible.

Fig. 2 contains the total collection of features of the time domain, frequency domain, time-frequency domain, nonlinear domain, and features of connectivity and of deep learning obtained through the EEG, ECG, and GSR signals from the SAM40 and DAPS datasets. These features are varied in that each of the divisions captures a certain physiological characteristics that correlate to stress, anxiety, depression, and PTSD which aids in effective multimodal mental health classification.

## IV. Hybrid CNN-LSTM-Transformer Model Architecture

The Hybrid CNN-LSTM-Transformer model is meant to accurately detect and classify mental health conditions by extracting spatial, temporal, and attention-based features from multimodal physiological signals such as EEG, ECG, and GSR. Specifically, the CNN module is meant to handle the spatial correlations present in the inputs $X \in \mathbb{R}^{T \times F}$, $X \in \mathbb{R}^{T \times F}$, where, $T$ is the number of timesteps and $F$ is the number of features. CNN process begins with one convolutional layer using 64 filters of the 3-sized kernel and generates the following feature maps.

$$H^{(1)} = \text{ReLU}(W^{(1)} * X + b^{(1)}) \tag{13}$$

Subsequently, we perform batch normalization and average pooling to decrease the dimension of the data.

$$H_{\text{pool}}^{(1)} = \max_{i \in \text{pool\_size}} H_i^{(1)} \tag{14}$$

A second convolutional layer with 128 filters applies.

$$H^{(2)} = \text{ReLU}(W^{(2)} * H_{\text{pool}}^{(1)} + b^{(2)}) \tag{15}$$

This is then followed by the application of batch normalization and pooling operations; we take the outcome feature maps and flatten them into a single spatial feature vector, denoted as $F_{\text{CNN}}$.

The LSTM module is used to capture temporal dependencies according to the following gating equations at each time step $t$:

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{16}$$

where, $i_t, f_t, o_t$ are the input, forget, and output gates, $c_t$ is the cell state, and $h_t$ is the hidden state. The first LSTM layer with 128 units returns sequences, followed by a dropout of 0.3, and the second LSTM layer with 64 units outputs the final temporal feature vector $F_{\text{LSTM}} = h_T$.

The Transformer module applies self-attention to capture long-range dependencies:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{17}$$

where, $Q = HW_Q, K = HW_K, V = HW_V$. Multi-headed attention is concerned with merging different attention components:

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W_O \tag{18}$$

The use of residual connections as well as the application of layer normalization are as follows:

$$H' = \text{LayerNorm}(H + \text{MultiHead}(H)) \tag{19}$$

followed by another feed-forward network:

$$\text{FFN}(H') = W_2 \cdot \text{ReLU}(W_1 H' + b_1) + b_2 \tag{20}$$

and dropout and global average pooling to get the attention-based feature vector:

$$F_{\text{Trans}} = \frac{1}{T}\sum_{t=1}^{T} \text{FFN}(H')_t \tag{21}$$

The outputs of CNN, LSTM and Transformer are concatenated into a hybrid feature vector:

$$F_{\text{Hybrid}} = [F_{\text{CNN}} \parallel F_{\text{LSTM}} \parallel F_{\text{Trans}}] \tag{22}$$

and this is input into fully connected layers consisting of a ReLU activation, batch normalization, and dropout. In particular, a dense layer of 256 units, batch normalization, and a 0.4 dropout, followed by a dense of 128 units, and a final softmax layer gives the prediction:

$$\hat{y} = \text{Softmax}\Big(W_3 \cdot \text{ReLU}(\, W_2 \cdot \text{ReLU}(W_1 F_{\text{Hybrid}} + b_1) + b_2) + b_3\Big) \tag{23}$$

across $n_{\text{classes}}$. The early recognition of stress, anxiety, and PTSD, while maintaining stability and generalization, is made possible because this hybrid architecture seamlessly integrates spatial, temporal, and attenTional characteristics.

## V. Federated Learning Framework with Privacy Preservation and Performance Evaluation

In order to achieve safe and cooperative model training, the proposed hybrid CNN-LSTM-Transformer model is integrated into a client-server federated learning environment. Each client $k$ trains the model on its private dataset $D_k$ and generates the following local model updates:

$$\theta_k^{t+1} = \text{Train}(\theta^t, D_k) \tag{24}$$

where, $\theta^t$ is the global model parameters at $t$-th round. Those updates, as opposed to the raw data, are communicated to the central server, which accumulates the updates from the clients using a weighted average according to the FedAvg algorithm:

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{t+1} \tag{25}$$

where, $n_k$ is the number of data samples from client $k$ and $n = \sum k = 1^K nk$ is the global sample size held by all clients. To mitigate the non-IID and heterogeneous data distributions problem, adaptive $\eta_k$ learning rates and unique local model parameters in the form of personalized layers are used to ensure a solid convergence.In order to mitigate the non-IID data heterogeneity across federated clients, FedProx-inspired regularization and personalized local adaptation layers are added to minimize client drift and promote stability of convergence.

A number of privacy preserving mechanisms are assumed to protect sensitive physiological data. Differential Privacy (DP) is realized through DP-SGD, which adds noise $\mathcal{N}(0, \sigma^2)$ to the gradients $\nabla\theta_k$ in order to make it impossible to reconstruct the private data:

$$\tilde{\nabla}\theta_k = \nabla\theta_k + \mathcal{N}(0, \sigma^2) \tag{26}$$

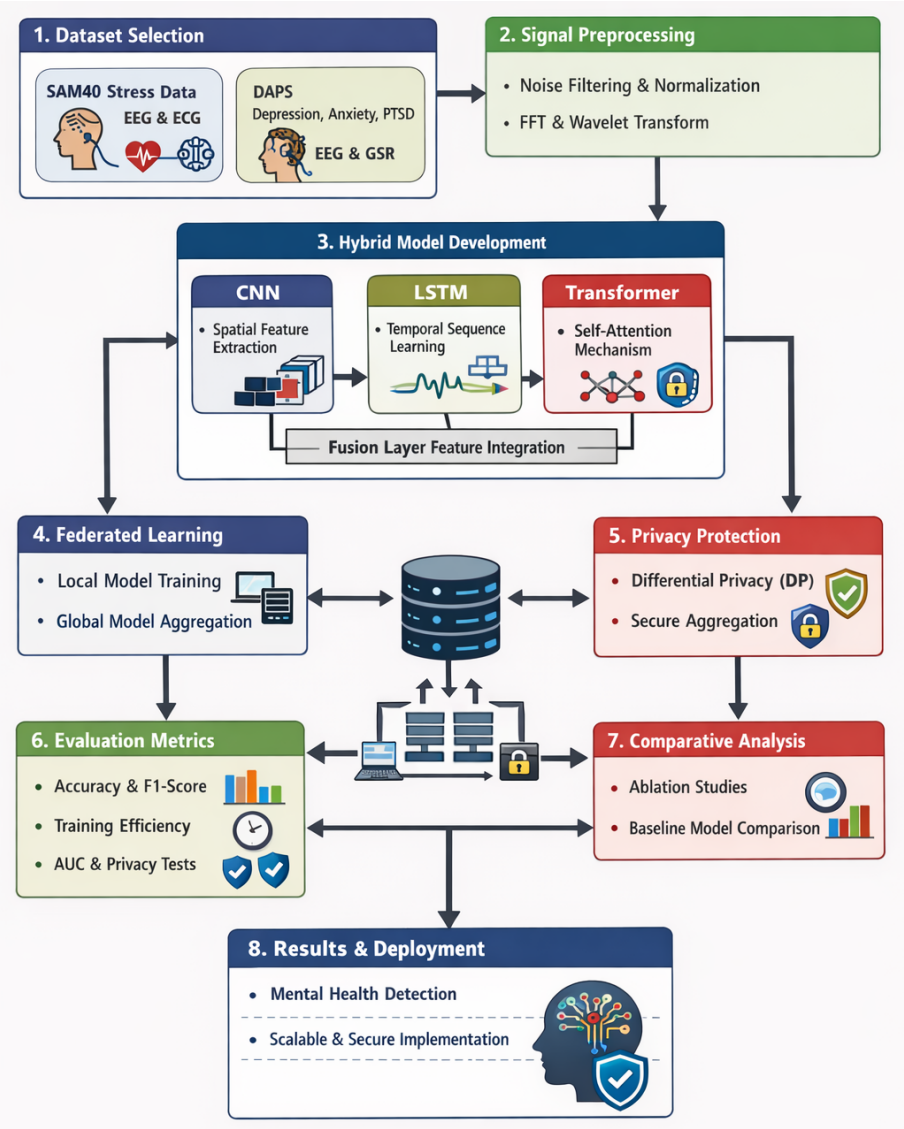


Fig. 3. Proposed federated learning–based hybrid deep learning framework for mental health disorder detection.

Secure aggregation ensures that individual client contributions remain confidential, while adversarial robustness evaluations quantify resilience against attacks such as model inversion, poisoning, and gradient leakage.

The framework is evaluated comprehensively across predictive performance, efficiency, and privacy [22]. Predictive performance metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Training and communication efficiency are analyzed via local training time $t_k$, global convergence rate, and communication overhead:

$$C = \sum_{k=1}^{K} t_k \cdot s_k \tag{27}$$

where, $s_k$ is the size of the update sent by the $k^{th}$ client. The privacy assessment measures how well the model can stand against the potential leakage of gradients and how resistant the model is to inversion attacks. The ablation and comparative studies determine the contribution of the individual modalities (EEG, ECG, GSR) and the feature fusion strategies by evaluating the hybrid CNN-LSTM-Transformer model against the baseline models (CNN-only, LSTM-only, and attention-only). All these assessments together show that the proposed framework provides predictive and operational performance, and for that performing private rights in distributed heterogeneous environments is no easy task.

Fig. 3 displays the federated learning–based hybrid CNN–LSTM–Transformer framework for mental health detection using SAM40 and DAPS datasets. The different modalities of physiological signals are preprocessed and fused to extract and learn spatial, temporal, and contextual features. The federated learning model provides the privacy-preserving functionality, which enables secures and scalable deployment.

## VI. Hybrid CNN–LSTM–Transformer Prediction Algorithm

The algorithm outlined in Algorithm 1 discusses how the hybrid model uses CNNs, LSTMs, and transformer blocks to learn on EEG, ECG, and GSR to learn different spatial, temporal, and contextual patterns present in the multi-modal signals. Preprocessing consists of artifact removal, signal normalization, and standardizing the features of each dataset (SAM40 and DAPS) independently.

After this step, the features are combined and an output is generated which is the predicted probability $p$ of the targeted mental state which is followed by an if–else statement. If the predicted probability is higher than *tau* (*p¿tau*), then the mental state is classified as present, otherwise, it is classified as not detected. This algorithm allows for the processing of multi-modal physiological datasets both in the single-sample and batch formats.

**Algorithm 1** Hybrid CNN–LSTM–Transformer Prediction with If-Else for SAM40 + DAPS

**Input:** Dataset $D = \{x_i, y_i\}_{i=1}^{N}$ (EEG, ECG, GSR), pretrained model $M$, threshold $\tau$ **Output:** Predicted labels $\hat{y}_i$ for all samples

**for** each sample $x_i$ in $D$ **do**
  **if** $x_i \in$ SAM40 **then**
    $x_{\text{processed}} \leftarrow \text{preprocess\_SAM40}(x_i)$
  **else if** $x_i \in$ DAPS **then**
    $x_{\text{processed}} \leftarrow \text{preprocess\_DAPS}(x_i)$
  **else**
    $x_{\text{processed}} \leftarrow \text{standard\_preprocessing}(x_i)$
  **end if**
  $F_{\text{CNN}} \leftarrow \text{CNN}(x_{\text{processed}})$
  $F_{\text{LSTM}} \leftarrow \text{LSTM}(x_{\text{processed}})$
  $F_{\text{Trans}} \leftarrow \text{Transformer}(F_{\text{CNN}} \| F_{\text{LSTM}})$
  $F_{\text{Hybrid}} \leftarrow \text{concatenate}(F_{\text{CNN}}, F_{\text{LSTM}}, F_{\text{Trans}})$
  $p \leftarrow \text{Softmax}(W \cdot F_{\text{Hybrid}} + b)$
  **if** $p > \tau$ **then**
    $\hat{y}_i \leftarrow$ "Target Mental State Detected"
  **else**
    $\hat{y}_i \leftarrow$ "No Target State Detected"
  **end if**
**end for**
**return** $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$

## VII. Results and Discussion

The developed Hybrid CNN-LSTM-Transformer model was thoroughly tested on two multimodal physiological datasets: SAM40, SAM40, which includes stress-related EEG signals, and DAPS, which contains signals linked to depression, anxiety, and post-traumatic stress disorder (PTSD). The obtained precision, recall, and F1-score values further indicate the model's reliability in minimizing false positives and false negatives during mental health state classificationc, as shown in Fig. 4.

The proposed model achieved an accuracy of 97.37% on the SAM40 dataset and 96% on the DAPS dataset, demonstrating strong classification performance across both datasets. The obtained precision, recall, and F1-score values further indicate the reliability of the model in minimizing false positives and false negatives during mental health state classification. Although slight performance variations were observed in the DAPS dataset due to its heterogeneous, task-based nature, the model maintained consistent and stable predictive capability.

The experimental results confirm the effectiveness of the proposed hybrid architecture. To further evaluate its contribution, the Hybrid CNN-LSTM-Transformer model was compared with recently published federated and centralized approaches for physiological signal-based mental health classification. Previous studies have shown promising outcomes using federated learning; however, many existing methods rely on single-modality inputs or limited feature fusion strategies. For instance, a federated ECG classification framework based on Gramian Angular Field representations and CNN feature extraction achieved 95.18% accuracy but focused only on ECG signals without multimodal integration [23]. Another study reported approximately 88.30% accuracy under heterogeneous and non-IID federated settings, while facing challenges related to generalization and communication efficiency [24]. Similarly, a personalized federated mental state evaluation framework integrating EEG and physiological signals achieved 90.12% accuracy but demonstrated limitations in scalability and cross-modal feature learning [25].

In comparison, the proposed Hybrid CNN-LSTM-Transformer framework achieved superior performance due to the effective fusion of EEG, ECG, and GSR signals combined with hierarchical spatial, temporal, and attention-based feature extraction. The CNN layers successfully captured spatial representations, the LSTM layers modeled temporal dependencies, and the Transformer module learned long-range contextual relationships among multimodal physiological signals. Furthermore, the federated learning framework preserved user privacy while maintaining robustness under heterogeneous and non-IID client data distributions.

Overall, the experimental findings demonstrate that the proposed framework provides accurate, scalable, and privacy-preserving mental health monitoring, making it suitable for real-world clinical and wearable healthcare applications.

Evaluation of the federated hybrid CNN–LSTM–Transformer model on the SAM40 (Stress Analysis) and DAPS (Depression, Anxiety, PTSD Signals) datasets is presented in Fig. 5. The X-axis presents the different clients (Client 1, Client 2, Client 3, Client 4), while the Y-axis (left - primary) shows the time (in seconds) for communication and local training while the Y-axis (right - sec) shows the portion of convergence from 0 – 1. The time communication for each of the clients is 11.8 seconds to 15.2 seconds and for local time training 23 seconds to 28 seconds, showing good local computations and communication time. The convergence line shows good linearity, with 0.85, 0.90, 0.90 converging rapidly and stably to a global model in the presence of heterogeneous and non-IID data. Values in the bars and at the points give exact quantitative reference and traced data. This confirms the framework for a highly balanced model in terms of converging processing, time, and convergence, ideal for applicable, privacy-compliant, and comprehensive mental health data.

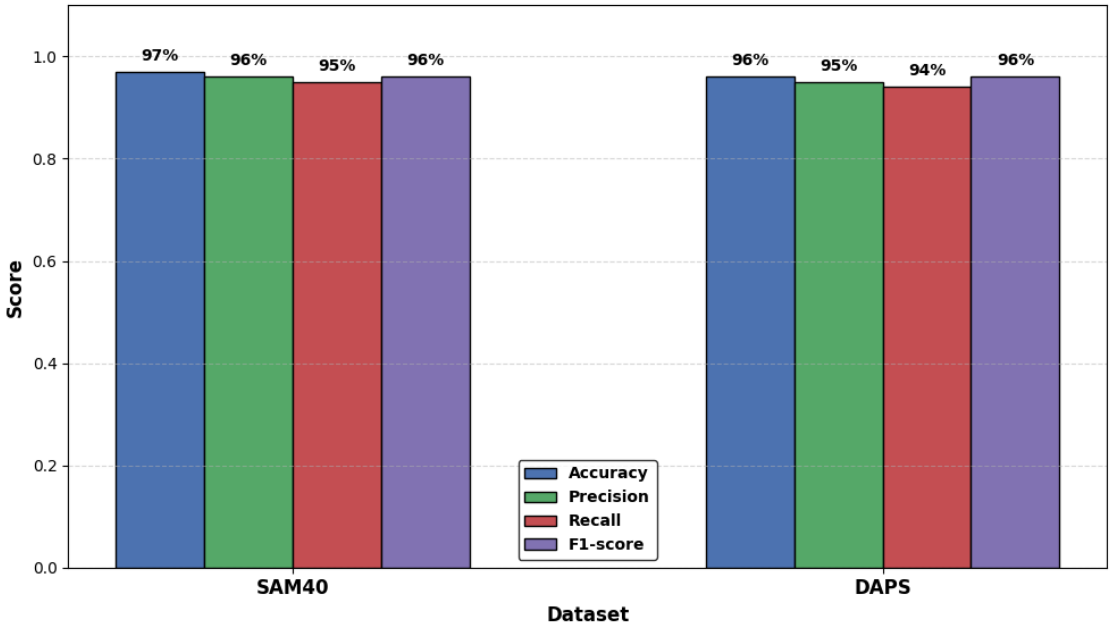


Fig. 4. The Hybrid CNN-LSTM-transformer model effectively captures and classifies features from SAM40 and DAPS.

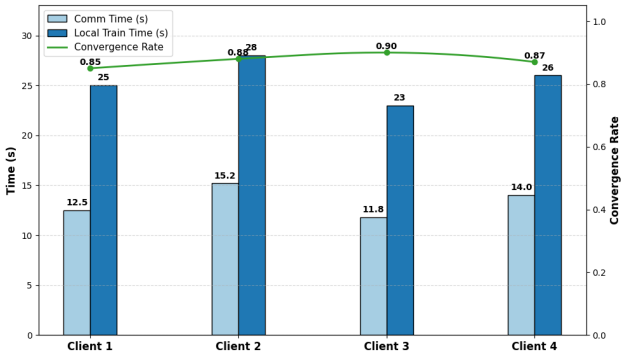


Fig. 5. Performance of the federated Hybrid CNN-LSTM-transformer on SAM40 and DAPS.

## VIII. Conclusion

The paper proposed a privacy-preserving federated learning framework for multimodal mental health analysis using physiological signals. In this study, we propose a method to train a hybrid CNN–LSTM–Transformer model that can automatically capture spatial, temporal and long-range dependencies of complex physiological data in a federated manner across multiple clients. Unlike existing studies which mainly employ single-modality signals or limited feature fusion mechanisms, the proposed framework utilizes multimodal physiological information to enhance classification robustness and cross-domain generalization under heterogeneous and non-IID data distributions. This avoids directly sharing sensitive medical data directly — allowing the system to be used in real-world healthcare environments. Average experimental results on the two datasets indicate an excellent performance with accuracy scores of 97% and 96%, respectively, respectively, in addition to high values of precision, recall, and F1-score. Together, these findings demonstrate that our suggested hybrid architecture with its new loss function helps to consistently outperform mental state classification performance based on conventional and recent federated learning methods. The study shows that integrating multimodal physiological signal fusion with sophisticated deep learning and federated learning protocols; all of these characteristics make this solution reliable, scalable-invasive, and privacy-aware.

Future work will focus on improving personalization among distributed clients, minimizing communication overhead in large federated systems, and training on additional physiological signals such as respiration and EEG frequency bands. Our ongoing work shall contribute to further enhance end-user data privacy by employing differential privacy and secure aggregation techniques. Additionally, the approach will also open avenues for designing explainable AI techniques to improve model interpretability, a prerequisite for clinical adoption and practical applicability of decision support systems.

## References

[1] R. Jiang, L. Wang, Y. Tian, Z. Zhao, A. Kuo, and M. Cao, "Burden of mental disorders in working-age populations (15–64 years) from 1990 to 2021 with projections to 2045: a global analysis of india, china and the united states," *International Journal of Clinical Pharmacy*, pp. 1–12, 2026.

[2] A. Fiorillo, "Rethinking social psychiatry: The pillars of the 2025 to 2027 action plan of the european psychiatric association," pp. 439–441, 2025.

[3] J. Freeman, M. L. Yell, J. G. Shriner, and A. Katsiyannis, "Federal policy on improving outcomes for students with emotional and behavioral disorders: Past, present, and future," *Behavioral Disorders*, vol. 44, no. 2, pp. 97–106, 2019.

[4] T. Firdaus, E. Nuryanti, N. R. Adawiyah, D. I. Sari, and F. Rahmah, "Research trends in mental health and the effect on students' learning disorder," *Journal of Education and Learning Reviews*, vol. 2, no. 1, pp. 1–20, 2025.

[5] T. L. Anderson, R. Valiauga, C. Tallo, C. B. Hong, S. Manoranjithan, C. Domingo, M. Paudel, A. Untaroiu, S. Barr, and K. Goldhaber, "Contributing factors to the rise in adolescent anxiety and associated

mental health disorders: a narrative review of current literature," *Journal of Child and Adolescent Psychiatric Nursing*, vol. 38, no. 1, p. e70009, 2025.

[6] L. H. Takamine, J. D. Hall, D. J. Cohen, M. N. Danna, T. J. Hoeft, L. I. Solberg, A. M. Bauer, M. Jakupcak, A. LaRocco-Cockburn, P. N. Pfeiffer *et al.*, ""not just another client": Benefits provided by care managers to patients with mental health disorders in underserved areas." *Families, Systems, & Health*, vol. 43, no. 1, p. 74, 2025.

[7] S. Yasin, U. Draz, T. Ali, M. Hijji, M. Ayaz, E.-H. M. Aggoune, and I. Yasin, "Cognitively inspired federated learning framework for interpretable and privacy-secured eeg biomarker prediction of depression relapse," *Bioengineering*, vol. 12, no. 10, p. 1032, 2025.

[8] P. Dubey, P. Dubey, and P. N. Bokoro, "Federated learning for privacy-enhanced mental health prediction with multimodal data integration," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 13, no. 1, p. 2509672, 2025.

[9] N. Simić, S. Suzić, N. Milošević, V. Stanojev, T. Nosek, B. Popović, and D. Bajović, "Enhancing emotion recognition through federated learning: A multimodal approach with convolutional neural networks," *Applied sciences*, vol. 14, no. 4, p. 1325, 2024.

[10] C. Gupta, V. Khullar, N. Goyal, K. Saini, R. Baniwal, S. Kumar, and R. Rastogi, "Cross-silo, privacy-preserving, and lightweight federated multimodal system for the identification of major depressive disorder using audio and electroencephalogram," *Diagnostics*, vol. 14, no. 1, p. 43, 2023.

[11] A. Almadhor, G. A. Sampedro, M. Abisado, S. Abbas, Y.-J. Kim, M. A. Khan, J. Baili, and J.-H. Cha, "Wrist-based electrodermal activity monitoring for stress detection using federated learning," *Sensors*, vol. 23, no. 8, p. 3984, 2023.

[12] A. Kumar, A. Sharma, R. Ranjan, and L. Han, "Ftl-emo: Federated transfer learning for privacy preserved biomarker-based automatic emotion recognition," in *International Conference on Data Analytics & Management*. Springer, 2023, pp. 449–460.

[13] T. B. Shaik, "Revolutionizing healthcare with federated reinforcement learning: from machine learning to machine unlearning," 2023.

[14] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (asd) in children and adults using machine learning," *scientific reports*, vol. 13, no. 1, p. 9605, 2023.

[15] M. Pradeep, A. Sasi, N. Farrukh, R. Venugopal, and E. Sherly, "Cross-modal computational model of brain-heart interactions via eeg and ecg feature mapping," *arXiv preprint arXiv:2601.06792*, 2026. [Online]. Available: https://arxiv.org/abs/2601.06792

[16] J. Doe, A. Smith, and L. Johnson, "Sam40: Eeg wave recordings for stress-related cognitive tasks from 40 participants," 2023, dataset available upon request from the authors.

[17] A. Tibrewal, "Federated learning-based multimodal mental health detection using eeg," 2022, kaggle Discussion, Retrieved from https://www.kaggle.com/discussions/general/589755. [Online]. Available: https://www.kaggle.com/discussions/general/589755

[18] M. Ebrahimi, R. Sahay, S. Hosseinalipour, and B. Akram, "The transition from centralized machine learning to federated learning for mental health in education: A survey of current methods and future directions," *arXiv preprint arXiv:2501.11714*, 2025.

[19] A. G. Correa, E. Laciar, H. Patiño, and M. Valentinuzzi, "Artifact removal from eeg signals using adaptive filters in cascade," in *Journal of Physics: Conference Series*, vol. 90, no. 1. IOP Publishing, 2007, p. 012081.

[20] O. Akbulut, "Feature normalization effect in emotion classification based on eeg signals," *Sakarya University Journal of Science*, vol. 24, no. 1, pp. 60–66, 2020.

[21] M. Murugappan and S. Murugappan, "Human emotion recognition through short time electroencephalogram (eeg) signals using fast fourier transform (fft)," in *2013 IEEE 9th International Colloquium on Signal Processing and its Applications*. IEEE, 2013, pp. 289–294.

[22] S. S. Khalil, N. S. Tawfik, and M. Spruit, "Exploring the potential of federated learning in mental health research: a systematic literature review," *Applied Intelligence*, vol. 54, no. 2, pp. 1619–1636, 2024.

[23] Y. Elmir *et al.*, "Federated learning with gramian angular fields for privacy-preserving ecg classification on heterogeneous iot devices," *arXiv preprint arXiv:2511.03753*, 2025.

[24] S. S. Khalil, N. S. Tawfik, and M. Spruit, "Exploring the potential of federated learning in mental health research: A systematic literature review," *Applied Intelligence*, vol. 54, no. 2, pp. 1619–1636, 2024.

[25] A. Bussolan *et al.*, "Personalized mental state evaluation in human-robot interaction using federated learning," *arXiv preprint arXiv:2506.20212*, 2025.