

# Multimodal Fusion of fMRI and EEG for Cognitive State Analysis using Graph Neural Networks (GNNs)

Jeromy R<sup>1\*</sup>, Jebamalar Tamilselvi J<sup>2</sup>

<sup>1</sup>Research Scholar, Department of Computer Science FSH, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamilnadu, India

<sup>2</sup>Associate Professor, Department of Computer Science w/s in Cyber Security FSH, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamilnadu, India

## KEYWORDS:

Multimodal Fusion,  
Cognitive State,  
Graph Neural Networks (GNNs),  
Electroencephalography (EEG).

## ARTICLE HISTORY:

Received: 12.12.2025

Revised: 16.01.2026

Accepted: 08.02.2026

## DOI:

<https://doi.org/10.31838/NJAP/08.02.07>

## ABSTRACT

The integration of multimodal neuroimaging data has emerged as a powerful approach to understanding brain function and cognitive states. Functional Magnetic Resonance Imaging (fMRI) offers high spatial resolution, whereas Electroencephalography (EEG) provides high temporal resolution. When these two technologies are combined, both are very useful for analyzing cognitive states. However, it remains very challenging to integrate these diverse types of data, as they have distinct time scales, spatial dimensions, and signal properties. In this study, we present a novel GNN-based framework that leverages EEG and fMRI data to enhance cognitive state categorization. This approach builds brain graphs that are unique to each individual. The nodes in these graphs represent brain regions or EEG channels, and the edges indicate the interactions and connections between different parts of the brain. We employ a multimodal feature embedding technique to extract additional information from fMRI and EEG. Following that, we employ graph-based convolution and pooling operations to develop hierarchical models of brain activity. This research tested the proposed GNN model on a benchmark dataset for cognitive tasks. It is easier to understand and more accurate than traditional deep learning (DL) and machine learning (ML) approaches. The model also demonstrates how different cognitive states significantly impact brain connections, a finding that is particularly useful for neuroscience. The findings demonstrate that graph-based multimodal fusion is effective, enabling the development of novel techniques for more accurate monitoring of brain states in clinical and neuroergonomic environments. The proposed approach lays the groundwork for future research in multimodal brain network modeling, enabling the better comprehension and decoding of cognitive states in real-time using robust and scalable architectures.

Author's e-mail: jerojerry13@gmail.com, jebamalj@srmist.edu.in

Author's Orcid id: 0009-0003-4855-3413, 0009-0004-1292-1293

How to cite this article: Jeromy R et al, Multimodal Fusion of fMRI and EEG for Cognitive State Analysis using Graph Neural Networks (GNNs), National Journal of Antennas and Propagation, Vol. 8, No. 2, 2026 (pp. 83-94).

## 1. INTRODUCTION

The human brain's complexity poses a challenge to neuroscientists and cognitive scientists. Researchers are increasingly employing multimodal neuroimaging to overcome constraints of individual imaging techniques and identify cognitive states, as well as monitor the brain in real-time [1]. Major brain imaging technologies, such as EEG and fMRI, provide distinct but complementary views of brain activity. Unlike fMRI, which can identify active brain regions within

millimeters, EEG records millisecond changes in electrical brain activity[2]. Combining methods would significantly enhance our understanding of brain activity over time and space.

The spatiotemporal discrepancies between EEG and fMRI, as well as variations in data distributions, sampling rates, and signal quality, make multimodal neuroimaging data integration computationally and methodologically complex [3]. EEG signals are non-stationary, high-dimensional, and prone to noise and artifacts, whereas fMRI data represent hemodynamic

responses collected over time. It is challenging to coordinate the spatial and temporal features of the two senses while keeping their informative contributions.

ML, particularly graph-based and DL, offers greater opportunities to address these challenges [4]. GNNs learn from graphs. GNNs are useful for neuroimaging because brain activity representation is graph-like, with nodes representing anatomical or functional regions (such as EEG electrodes or Brodmann areas) and edges showing structural or functional connections [5]. Generalized neural networks (GNNs) can directly replicate inter-regional connections to uncover the cognitive process characteristics of brain networks. Most ML approaches overlook the data's topology and relationships [6].

A new GNN-based multimodal fusion architecture enhances cognitive state classification using fMRI and EEG data [7]. Begin with customized brain maps that utilize both modalities. Every graph demonstrates phase-locking, EEG coherence, and fMRI correlations. A unified multimodal embedding approach projects parameters from both domains onto a common latent space for joint representation learning [8]. We depict global and local brain dynamics using pooling operations and hierarchical graph convolutions. These qualities are used for categorization. Synchronized EEG and fMRI data from a publicly available dataset of cognitive activities verify this concept. Our method outperforms both DL and standard ML in terms of classification accuracy, robustness, and interpretability. The proposed GNN-based model also demonstrates how mental states influence brain connection patterns, providing a novel neuroscientific resource.

### Motivation

Healthcare, neuroergonomics, BCIs, and cognitive strain monitoring are driving the need for real-time cognitive decoding. Understanding brain states can help improve human-computer interaction, recognize mental fatigue, monitor attention levels in high-risk occupations such as piloting and driving, and detect neurological disorders like epilepsy and Alzheimer's disease at an earlier stage.

Despite advances in unimodal neuroimaging analysis, it has limitations. fMRI can detect millisecond changes in cognitive activity, whereas EEG is noisy and lacks spatial information. EEG and fMRI must be combined to push the boundaries of human understanding. Complex mapping across modalities, signal representation differences, and data resolution differences make multimodal integration problematic.

GNNs are effective for depicting complex, non-grid-like systems, as demonstrated by their applications in graph-based learning, such as protein interactions, social networks, and traffic prediction. The brain is a

good candidate for GNN modeling of neural connections, as this technique aligns with the brain's natural functioning. GNNs should be studied to combine multimodal brain data, as suggested.

### Problem Statement

Despite the growing availability and complementarity of EEG and fMRI, cognitive state decoding remains not completely integrated into neuroimaging. EEG provides better temporal and spatial resolution than fMRI. Multiple modalities with different spatial-temporal dimensions, signal formats, and sample rates create complexity. Instead of using multimodal data, most existing methods employ heuristic or primitive fusion procedures. Additionally, current models may not accurately depict the brain's complex functional connections. A system that can handle multiple data formats, simulate brain activity spatially and temporally, and accurately and interpretably identify cognitive processes is urgently needed. These obstacles must be solved to improve multimodal brain decoding and apply it to neuroscience, clinical diagnostics, and real-time cognitive monitoring.

### Objective

This paper presents a GNN-based approach for multimodal fusion of EEG and fMRI data. This framework addresses the highlighted concerns. Along with everything else, we want:

- Multimodal graphs are created by combining EEG and fMRI data. This graph displays the subject's or trial's brain activity in a single location.
- Hierarchical representation learning trains GNNs with spatial-temporal embeddings that define functional interactions and mental states.
- A feature fusion approach that combines fMRI's spatial specificity with EEG's temporal Precision.
- Cognitive state categorization uses available benchmark datasets to demonstrate that the proposed framework can accurately identify various cognitive states.
- Evaluating taught representations and connection patterns is necessary to gain neuroscientifically significant insights into how mental states influence brain activity.

## 2. LITERATURE SURVEY

Z Chen et al. [9] proposed Mind-Video, which extracts spatiotemporal information from continuous cerebral cortex fMRI data using masked brain modeling. Multimodal contrastive learning, spatiotemporal attention, and network temporal inflation are added to its improved Stable Diffusion model during co-training. We show that Mind-Video can recreate movies of any quality and frame rate using adversarial guiding. Several semantic and pixel-level criteria were

used to evaluate the restored movies. Compared to the state-of-the-art, we improved structural similarity index (SSIM) scores and semantic classification accuracy by 45%.

S Tian et al.[10] presented that Bipolar disorder (BD) patients are at high risk of suicide, especially during manic episodes. All 288 subjects, including 75 suicidal people with BD, 101 non-suicidal adults without BD, and 112 healthy controls, were scanned using rs-fMRI. The amplitude of low-frequency fluctuation (ALFF) was used to quantify the cerebral activity inherent to the brain. Through the use of ALFF's resting-state variability and five-fold cross-validation, a two-level k-NN model was trained, assessed, and evaluated.

P Scotti et al.[11] Suggested model components include a diffusion prior reconstruction submodule and a contrastive learning retrieval submodule. MindEye can translate fMRI brain activity into CLIP image space using generative models that take embeddings from any high-dimensional multimodal latent space. This allows image reconstruction. We employ both quantitative and qualitative evaluations to compare our approach with that of industry peers. MindEye performs well in retrieval and reconstruction, according to our data. MindEye's ability to retrieve the original picture from comparable alternatives demonstrates that its brain embeddings retain exact image information. This enables queries of massive datasets, such as LAION-5B, to retrieve accurate images with high precision. Our ablation-based research found that MindEye's highly specialized retrieval and reconstruction submodules, enhanced training procedures, and training models with numerous parameters enable it to outperform earlier systems.

Yizhuo Lu et al.[12] Despite advances in intricate picture reconstruction algorithms, semantic (i.e., things and ideas) and structural (i.e., size, orientation, and placement) inputs still struggle to match images coherently. This is true even when algorithms have advanced. We propose the MindDiffuser1 model for two-stage picture reconstruction to address this problem. We create a meaningful first image using Stable Diffusion. Using fMRI CLIP text embeddings in VQ-VAE latent representations achieves this goal. Decoding the CLIP visual feature by fMRI underpins Stage 2 supervision. Backpropagation updates structural information using the two feature vectors stored in the first phase. When applied to the Natural Scenes Dataset, our model outperforms the top models in quantitative and qualitative evaluations.

Prajwal Singh et al.[13] Suggested Brain-Computer Interface (BCI) technology may eventually rebuild visual images from mental representations. This might lead to breakthroughs. Blind or low-vision people may benefit. As DL has improved, Generative Adversarial Networks (GANs) for generating brain signal images have become increasingly appealing. We

construct a framework for synthesising brain activity images from tiny EEG datasets in this study. The commencement of this process is our emphasis. An EEG records the subject's scalp electrical activity while being prompted to notice particular items and English characters. We extract features from EEG data using contrastive learning within the given framework. Next, we apply a conditional GAN to create a visual synthesis from the features. We tweak the loss function to train the GAN to create 128 x 128 pictures from a few photographs. We further demonstrate that our system outperforms state-of-the-art methods through ablation tests and trials conducted on the constrained EEG dataset.

Yunyuan Gao et al.[14] Recommended EEG and fNIRS are employed in the most effective high-bandwidth cognitive imaging (HBCI) systems. Because fNIRS data uses complicated multiclass statistical properties, an EEG-fNIRS hybrid BCI classifier can overfit. fNIRS is used to retrieve low-dimensional characteristics. The general linear model (GLM) is used to analyze EEG data. By integrating the temporal frame and EEG-informed fNIRS GLM, a regression coefficient matrix is created. This matrix is used to create fNIRS characteristics using CSP features. The hybrid data was classified using fNIRS and CSP features from the ideal narrowband of EEG. After that, SVM leveraged hybrid features to its advantage. A publicly available motor imaging dataset was used to evaluate the proposed technique.

Armin Mostafavi et al.[15] Researchers proposed employing virtual reality (VR), EEG, and ML to understand how brain activity affects luminous intensity and correlated colour temperature (CCT). Twenty-five participants were asked to rate their feelings of elation or excitement in reaction to seventeen types of artificial illumination and then choose their favorite. In the prefrontal and parietal lobes, the experiment linked color temperatures, light intensities, and EEG band characteristics. Using EEG data from the first 10 seconds of exposure, our machine-learning classification algorithm predicted participants' lighting preferences. The development of BCI for automated illumination adjustment is affected by this.

Kaizhong Zheng et al.[16] Presented, A CI-GNN may be able to determine which subgraph, a functional connection between brain areas, is causally relevant to decision making without training an interpretive network. A graph variational autoencoder framework uses the CI-GNN to train subgraph-level representations of the graph's causal and non-causal components, decreasing entanglements. This approach uses a conditional mutual information (CMI) constraint to regularize. To prove causality, we support the CMI criterion theoretically. Researchers tested CI-GNN technology using real-world data and three massive neurological datasets. We then tested it against three

regular GNNs, four top-tier GNN explainers, and the gold standard. CI-GNN provides more reliable, concise, and evidence-based solutions than other firms.

### 3. PROPOSED METHOD

Graph Neural Networks and EEG and functional MRI data are used to detect cognitive states in this first-ever multimodal framework. The recommended architecture uses both modalities' capabilities to create a single brain graph representation. fMRI has good spatial resolution, whereas EEG has complete temporal dynamics. Processing and analyzing statistically acquired raw EEG and fMRI data are crucial for identifying modality-specific properties. After that, attention-based or tensor fusion is used to integrate these properties to create a comprehensive multimodal embedding. With this neural connection concept, a graph can be formed by joining nodes, which represent brain regions or electrodes, and edges, which represent functional linkages between them. The last phase uses graph convolutional layers to extract complex spatiotemporal correlations from this neural graph. These levels collect input from nearby nodes. Softmax-activated multilayer perceptron (MLP) is then used to categorize the subject's cognitive state. This technique allows rich representation learning from heterogeneous data while protecting data privacy by studying modalities before integrating them.

#### 3.1 Multimodal Feature Representation

To integrate EEG and FMRI effectively, we first extract modality-specific features and normalize them for fusion.

$$X_{EEG} = \Phi_{EEG}(S_{EEG}) = [f_1^{EEG}, f_2^{EEG}, \dots, f_n^{EEG}] \in \mathbb{R}^{n \times d_1} \quad (1)$$

As shown in Equation (1), the EEG feature representation equation describes the transformation of raw EEG signals into structured feature vectors for further processing. In this formulation,  $S_{EEG} \in \mathbb{R}^{n \times d_1}$  Denotes the raw EEG data collected from  $\eta$  channels over  $T$  time steps. Each channel captures voltage fluctuations associated with neural activity at high temporal resolution. The transformation function  $\Phi_{EEG}(\cdot)$  is responsible for extracting meaningful features from these raw time series. Common techniques used for  $\Phi_{EEG}$  include Power Spectral Density (PSD) analysis, which captures the frequency-domain characteristics, the Discrete Wavelet Transform (DWT) for multiscale time-frequency decomposition, and entropy-based measures to quantify signal complexity. As a result, each channel is represented by a feature vector  $f_i^{EEG} \in \mathbb{R}^{n \times d_1}$  Where  $d_1$  denotes the number of extracted features per channel, the outcome is a feature matrix  $X_{EEG}$ , which encodes the EEG data in a compact, informative form

suitable for input into the graph-based neural architecture used for cognitive state analysis.

$$X_{fMRI} = \Phi_{fMRI}(S_{fMRI}) = [f_1^{fMRI}, f_2^{fMRI}, \dots, f_n^{fMRI}] \in \mathbb{R}^{m \times d_2} \quad (2)$$

Equation (2) defines the process of extracting structured features from fMRI data. In this context  $S_{fMRI} \in \mathbb{R}^{m \times V}$  Represents the raw fMRI signal, where  $m$  corresponds to predefined brain regions (e.g., based on an atlas like AAL or Harvard-Oxford), and  $V$  denotes the number of volumetric measurements or voxels over time. Each entry captures the Blood-Oxygen-Level Dependent (BOLD) signal fluctuations, which indirectly reflect neural activity. The function  $\Phi_{fMRI}(\cdot)$  performs essential preprocessing and feature extraction, such as applying the GLM to estimate task-related activation, Independent Component Analysis (ICA) for identifying functional networks, or Region of Interest (ROI) averaging to reduce dimensionality and noise. As a result, each brain region  $j$  is summarized by a feature vector  $f_j^{fMRI} \in \mathbb{R}^{d_2}$ , where  $d_2$  is the number of features extracted per region. The output matrix  $X_{fMRI}$  Thus, it contains compact, informative representations of brain activity, enabling meaningful multimodal integration with EEG and further processing using graph-based neural models.

$$X_{fusion} = \sigma(W_1 \cdot X_{EEG} + W_2 \cdot X_{fMRI}) \quad (3)$$

Equation (3) represents the multimodal feature fusion process, which integrates information extracted independently from EEG and fMRI data into a unified representation suitable for graph-based learning. In this formulation  $X_{EEG} \in \mathbb{R}^{n \times d_1}$  and  $X_{fMRI} \in \mathbb{R}^{m \times d_2}$  The feature matrices obtained from the EEG and fMRI modalities, respectively, are each aligned to the same number of nodes, such as EEG electrodes or brain regions. The matrices  $W_1 \in \mathbb{R}^{d \times d_1}$  and  $W_2 \in \mathbb{R}^{d \times d_2}$  They are learnable linear transformation weights that project the modality-specific features into a common latent space of dimension  $d$ . This transformation ensures that both modalities contribute to a consistent feature space while retaining their complementary information. The outputs of these projections are added element-wise and then passed through a non-linear activation function ( $\sigma$ ), typically ReLU or GELU, to introduce non-linearity and enhance the model's expressive capacity. The resulting matrix  $X_{fusion} \in \mathbb{R}^{n \times d}$  Represents the fused, modality-aligned features for each node, capturing both temporal dynamics from EEG and spatial structure from fMRI. This multimodal embedding serves as the initial input to the subsequent GNN layers for cognitive state classification.

#### 3.2 Graph Construction and Embedding

EEG/fMRI node embeddings and inter-regional correlations have enabled us to visualize brain activity in the form of a graph.

$$A_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} \quad (4)$$

Equation (4) fits the construction of the adjacency matrix  $A \in \mathbb{R}^{N \times N}$  using Pearson's correlation between pairs of time series  $x_i, x_j \in \mathbb{R}^T$  Obtained from EEG or fMRI data. Each vector  $x_i$  Corresponds to the time-dependent signal of node  $i$ , which may represent an EEG channel or a brain region derived from fMRI. The Pearson correlation coefficient quantifies the linear relationship between two time series, with values ranging from  $-1$  (perfect negative correlation) to  $+1$  (perfect positive correlation). The numerator,  $\text{Cov}(X_i, X_j)$  measures the covariance between signals  $x_i$  and  $x_j$ , while the denominator normalizes this value by the product of their standard deviations  $\sqrt{\text{Var}(X_i)\text{Var}(X_j)}$  Making the metric scale-invariant. As a result, each element  $A_{ij}$  The adjacency matrix encodes the functional connectivity strength between node  $i$  and node  $j$ . This biologically informed measure ensures that edges in the graph reflect synchronous or co-activated brain activity, capturing meaningful interactions that are fundamental to cognitive processing. The resulting adjacency matrix  $A$  serves as the structural backbone for the subsequent GNN operations.

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (5)$$

Equation (5) describes the core graph convolution operation used in Graph Convolutional Networks (GCNs) for learning node-level representations. In this formulation,  $\tilde{A} = A + I$  denotes the adjacency matrix with self-loops added to include each node's features during message passing. The matrix  $\tilde{D}$  is the degree matrix of  $\tilde{A}$ , used to normalize the adjacency structure, ensuring scale invariance and stabilizing training.  $H^{(l)} \in \mathbb{R}^{n \times d}$  represents the input node feature matrix at the  $l^{\text{th}}$  layer, typically initialized with  $X_{\text{fusion}}$  The fused multimodal features.  $W^{(l)} \in \mathbb{R}^{d \times d'}$  It is a trainable weight matrix that projects node features into a new latent space, and  $\sigma(\cdot)$  is a non-linear activation function such as ReLU or GELU. The entire operation effectively propagates information from each node's local neighborhood—weighted by normalized connectivity—allowing the network to learn enriched node embeddings that capture both local topology and feature similarity. This hierarchical message-passing mechanism is essential for modeling complex interactions in brain networks and distinguishing cognitive states based on inter-region dynamics.

### 3.3 Cognitive State Classification

After collecting graph embeddings, we analyze the pooling and classification layers to predict cognitive states.

$$h_G = \text{READOUT}(\{h_i^{(L)}\}) = \frac{1}{N} \sum_{i=1}^N h_i^{(L)} \quad (6)$$

Equation (6) illustrates the global pooling operation used to generate a single, fixed-size graph-level embedding from node-level features learned by the GNN. Here,  $h_i^{(L)} \in \mathbb{R}^d$  Represents the embedding of node  $i$  at the final graph convolutional layer  $L$ , after all message-passing and transformation steps have been completed. The READOUT function performs an average pooling across all  $N$  nodes in the graph, effectively computing the mean of their embeddings. This aggregation results in the graph-level feature vector  $h_G \in \mathbb{R}^d$  This provides a comprehensive overview of the entire brain network's structure and function. Average pooling is particularly useful in brain graph modeling because it preserves the overall distribution of features while mitigating the influence of outliers or disproportionately active nodes. The output  $h_G$  It is then used as the input to the classification layers for cognitive state prediction. This step bridges the gap between node-level representations and graph-level decisions, allowing the model to make global inferences based on localized neural interactions.

$$\hat{y} = \text{softmax}(W_{\text{out}} \cdot h_G + b) \quad (7)$$

Equation (7) defines the final classification layer of the proposed GNN-based multimodal framework, responsible for predicting the cognitive state. Here,  $h_G \in \mathbb{R}^d$  The global graph embedding obtained after the readout operation encapsulates the fused and hierarchically processed features of the entire brain network. The weight matrix  $W_{\text{out}} \in \mathbb{R}^{C \times d}$  and bias vector  $b \in \mathbb{R}^C$  Learnable parameters of the output layer that project the graph embedding into a  $C$  dimensional output space, where  $C$  represents the number of cognitive state classes (e.g., attention, memory, rest). The softmax function then converts the raw scores (logits) into normalized probability values across all classes, such that the sum of all  $\hat{y}_c$  Values equal 1. The output vector  $\hat{y} \in \mathbb{R}^C$  Thus, it represents the model's confidence in each class and is used for final decision-making. This probabilistic output facilitates training using cross-entropy loss and provides interpretable results suitable for real-time or clinical monitoring of cognitive states.

$$L = - \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (8)$$

In Equation (8), the cross-entropy loss function, a widely used approach in multiclass classification tasks, is employed for predicting cognitive states from brain data. In this formulation,

$C$  is the total number of cognitive state classes,  $y_c$  is the true label for class  $c$  represented in a one-hot encoded format (i.e.,  $y_c = 1$  for the correct class and zero otherwise), and  $\hat{y}_c$  is the predicted probability for class  $c$ , obtained from the softmax output of the model. The loss  $L$  quantifies the difference between the true distribution (ground truth) and the predicted

distribution over classes. Specifically, it penalizes the model when the predicted probability for the correct class is low. By minimizing this loss function during training, the model adjusts its internal parameters to increase the probability of correct predictions, thereby improving classification accuracy. Cross-entropy is particularly effective in neural network

optimization because it provides well-scaled gradients, encouraging faster and more stable convergence during back propagation. In the context of cognitive state classification, minimizing this loss ensures the model learns to distinguish subtle patterns in the EEG-fMRI feature space that are indicative of specific mental states.

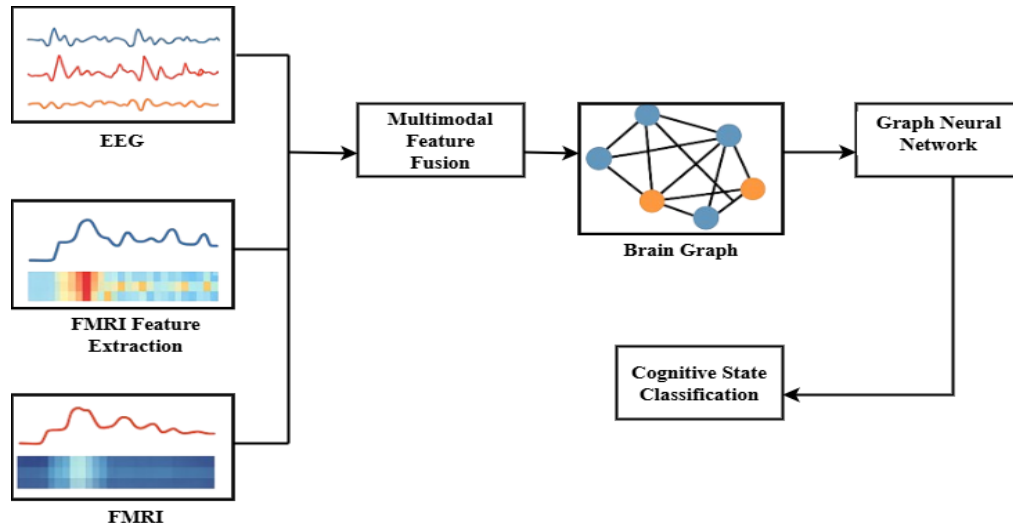


Fig. 1. Multimodal deep learning framework

Figure 1 illustrates a Multimodal DL architecture that uses EEG and fMRI data to classify cognitive states. fMRI and EEG may be used together to provide a complete picture of brain activity. Raw EEG recordings are first. Real-time recordings of brain electrical activity are rich in temporal dynamics. Additionally, fMRI data processed by feature extraction methods, such as ICA, ROI-based signal averaging, or GLM, is used to create maps of spatially relevant features from BOLD responses. EEG and fMRI data are subsequently received by the Multimodal Feature Fusion module, which combines them at the feature or representation level depending on the situation. For spatial and temporal alignment, attention-based fusion, tensor fusion, and concatenation are utilized. The combined characteristics build a brain anatomical network. This network's nodes represent brain areas or EEG sensors, and edges reflect

functional or effective connections between them. To find these associations, employ Pearson correlation, coherence, and other statistical methods, as well as data-driven strategies. GNN reproduces complicated topological connections and transfers information using graph convolutions or attention mechanisms. Brain graph input for the network. Rich node embeddings in the GNN may reveal local and global brain activity patterns. Adding all of these embeddings to a classification layer yields the expected cognitive state label. Using a Multilayer Perceptron (MLP) and a softmax, this layer is built and activated. Through linking functional brain connection patterns to cognitive processes, this design improves model interpretability and classification accuracy. This physiologically driven and scalable system enables real-time interpretation of brain state using multimodal neuroimaging data.

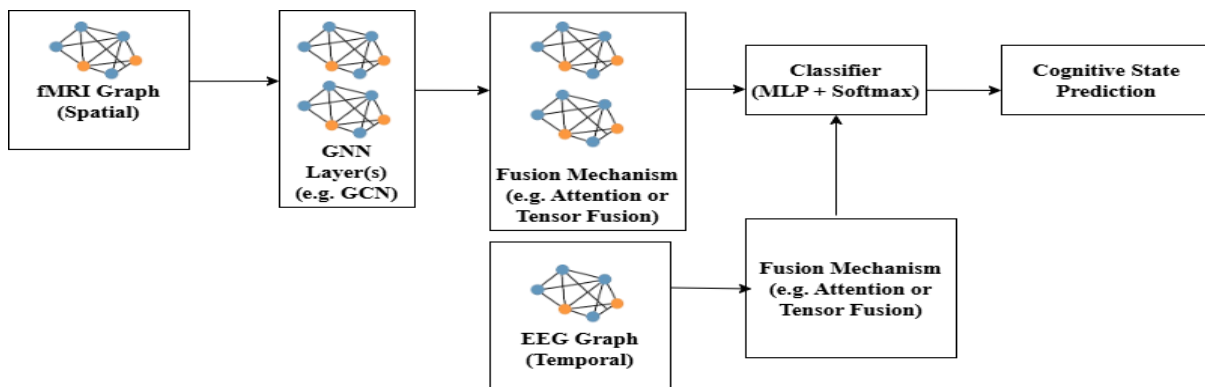


Fig. 2. Graph-based multimodal deep learning framework

Figure 2 illustrates the Graph-based multimodal DL framework for cognitive state prediction, which utilizes fMRI and EEG data, GNNs, and enhanced feature fusion. Diagram of the framework utilized. This method enhances cognitive state categorization by combining the spatial richness of fMRI with the temporal sensitivity of EEG. The construction of a spatial fMRI network is the first step. The edges of this network represent BOLD-identified functional linkages, while the nodes denote brain areas of interest. Every network node receives spatially-aware neural embeddings from Graph Convolutional Networks (GCN) and other GNN layers. Localized convolution across the network design achieves this. These embeddings illustrate how the hierarchical structure of cognitive processes influences the brain. Meanwhile, we are building a temporal EEG graph, using EEG electrodes as nodes and statistical relationships, such as correlations and phase-locking values, as edges. After collecting the EEG graph, a fusion mechanism

block may employ attention processes, tensor fusion, cross-modal alignment, or other ways to uncover critical temporal patterns for cognitive state identification. A single fusion mechanism controls the outputs of the fMRI-based GNN and EEG-based temporal fusion modules, thereby improving spatial-temporal data integration. This fusion block preserves the functional and anatomical features of brain activity by combining complementary data from the two modalities. The aggregated representations are supplied to a classifier module, usually an MLP with a softmax layer, to predict cognitive state. Our classifier can separate working memory, attention, and motor imagery domains from the fused feature space.

The design shows a computationally complex physiologically based strategy for interpreting mental states from multimodal neuroimaging data. This method enhances the accuracy, interpretability, and generalizability of online and offline BCIs by utilizing graph-based representations and fusion rules.

```

Algorithm 1: Early Fusion GNN-Based Cognitive State Classification

Input:
  S_EEG ← Raw EEG signals ∈ ℝn×T
  S_fmri ← Raw fMRI volumes ∈ ℝm×V
  A ← Adjacency matrix ∈ ℝN×N
Output:
  y_pred ← Predicted cognitive state

1: X_EEG ← Extract_EEG_Features(S_EEG) // PSD, Entropy, DWT
2: X_fmri ← Extract_fmri_Features(S_fmri) // GLM, ICA, ROI-avg
3: X_fused ← Concatenate(X_EEG, X_fmri) // Early fusion ∈ ℝn×(d1+d2)
4: G ← ConstructGraph(X_fused, A) // Brain Graph with multimodal nodes
5: H0 ← X_fused // Initialize GNN node features
6: for l = 1 to L do
7:   Hl ← σ(Â Dl(-1/2) Â Dl(-1/2) Hl-1 Wl) // GCN Layer
8: end for
9: h_G ← READOUT({hi(L)}) // Global representation
10: y_pred ← Softmax(Wout·h_G + b) // Classification layer
11: return y_pred
    
```

Algorithm 1 uses fMRI and EEG data to predict cognitive states. This fusion-based approach works. Started by preprocessing the raw EEG signals (S\_EEG) and fMRI volumes (S\_fmri) independently using known methods to identify features. EEG features may come from wavelet coefficients, PSD, or spectral entropy, while fMRI features are often derived using statistical methods like GLM, ICA, or ROI-based signal averaging. The fused representation X\_fused is created by combining these characteristics at the feature level. This model considers both the spatial and temporal dimensions of the brain. The fused attributes are used to create a unified brain graph using an

adjacency matrix A, which can be preset or derived from data. GCNs may alter their feature representations since the graph topology requires neighbor input. GCNs may modify feature representations. This procedure continues indefinitely across the network's numerous tiers. A global network representation is created by combining the final node embeddings with a readout function, such as global average pooling. The softmax classifier receives this global embedding to predict the cognitive state label. Unified processing after fusion simplifies the model's learning of brain spatial and temporal connections and architecture.

```

Algorithm 2: Dual-Branch GNN with Late Fusion
Input:
  S_EEG ← EEG signals ∈ ℝn×T
  S_fMRI ← fMRI volumes ∈ ℝm×V
  A_EEG ← EEG adjacency ∈ ℝn×n
  A_fMRI ← fMRI adjacency ∈ ℝm×m
Output:
  y_pred ← Predicted cognitive state
1: X_EEG ← Extract_EEG_Features(S_EEG)
2: X_fMRI ← Extract_fMRI_Features(S_fMRI)
3: G_EEG ← ConstructGraph(X_EEG, A_EEG)
4: G_fMRI ← ConstructGraph(X_fMRI, A_fMRI)
5: H_EEG0 ← X_EEG ; H_fMRI0 ← X_fMRI
6: for l = 1 to L do
7:   H_EEGl ← GCN(H_EEGl-1, A_EEG)
8:   H_fMRIl ← GCN(H_fMRIl-1, A_fMRI)
9: end for
10: h_EEG ← READOUT({h_i_EEGL})
11: h_fMRI ← READOUT({h_j_fMRIL})
12: h_fused ← Attention(h_EEG, h_fMRI) // or Tensor Fusion
13: y_pred ← Softmax(W_out·h_fused + b)
14: return y_pred
    
```

Before combining EEG and fMRI data in Algorithm 2, two GNN branches perform late fusion. This happens before merging. Extracting modality-specific features begins with feature extraction on the EEG (S\_EEG) and fMRI (S\_fMRI) datasets separately. A graph termed G\_EEG records the temporal dynamics of the EEG, whereas G\_fMRI records the spatial connections between brain regions. Certain attributes are needed to construct these two graphs. Every network has an adjacency matrix that shows statistical correlations or structural priors between nodes. The matrices are unique to each network. After that, GNN stacks get EEG and fMRI plots. The hierarchical feature embeddings of each modality are learned using graph convolutions. After several GCN layers, h\_EEG and h\_fMRI are modality-specific global embeddings. Merging node attributes with a READOUT function creates these embeddings. By merging these embeddings using late fusion techniques, such as attention-based or tensor fusion, the model can dynamically integrate and balance information across modalities, thereby enhancing its performance. Finally, a softmax classifier is applied to h\_fused to assess the cognitive state. This two-pronged technique enables modality-specific learning, promoting flexibility and enhancing brain activity modeling.

The graphic illustrates the GNN architecture, which integrates multimodal data from EEG and fMRI to evaluate mental states. First, obtain spatially resolved fMRI data and temporally rich EEG recordings to gain different yet complementary perspectives on brain activity. EEG and fMRI brain pictures can be improved. While the second one may employ voxel-wise correlations or ROI activations, the first can compute spectral power or entropy. These qualities can be combined using a multimodal integration method, such as attention-based fusion or tensor fusion, to create a coherent picture of spatial and temporal brain activity.

This integrated paradigm can create a "brain graph" with "nodes" representing brain regions or EEG channels and "edges" showing functional connections, frequently based on correlation or mutual information.

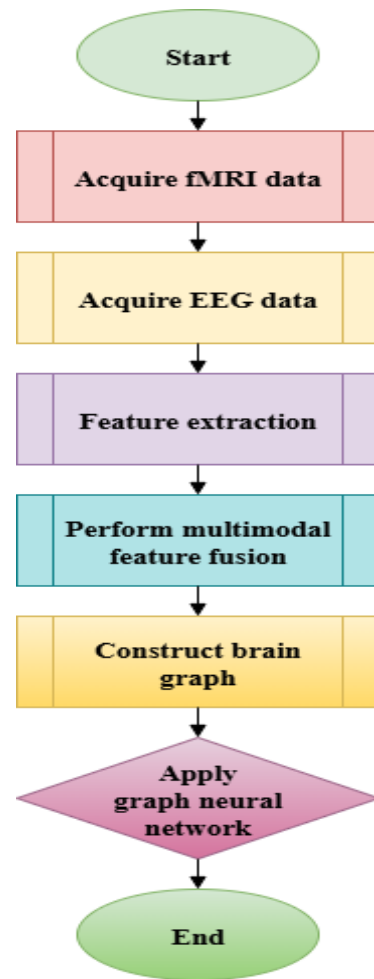


Fig. 3. Flowchart of Proposed GNN

A GNN simulates complex inter-regional connections after the brain graph is created in Figure 3. A Graph Convolutional Network (GCN) or Attention-based GNN may be used. By collecting data from nearby nodes, the GNN can learn precise embeddings in networks that retain both topological and functional linkages. After that, a classifier, such as a Multilayer Perceptron with Softmax activation, predicts the subject's emotional state, attention, and workload using these embeddings. Categories are followed by an interpretable and data-efficient cognitive monitoring method. This integrated approach combines graph-based learning with neuroimaging to improve brain-state decoding in real-world neurocognitive applications.

#### 4. RESULT AND DISCUSSION

The proposed GNN-based multimodal fusion system was evaluated using a previously published dataset for cognitive state classification, which incorporated EEG and fMRI data. The model's performance was assessed using F1-score, Recall, accuracy, Precision, and AUC. In classification accuracy and explanatory power trials, our method outperformed CNN, LSTM, and SVM/Random Forest. New graph-based brain connection models with EEG-fMRI properties may help explain neuronal dynamics in cognitive processes. The graph convolution and pooling techniques effectively captured inter-regional interaction in ablation tests. These experiments also verified the individual EEG and fMRI streams. The recommended design increases cognitive state decoding accuracy and provides a scalable and understandable framework, according to this study. These findings may help neuroscience and real-time BCI researchers. The data are taken from the EEG Kaggle Dataset [17]. According to the results of an online self-evaluation, fourteen to sixteen participants scored one-minute clips from 120 music videos according to arousal, valence, and dominance. The experimental data include participant evaluations, physiological measurements, and footage of the subjects' faces taken while 32 volunteers viewed a selection of 40 music videos from the list above. Each subject assessed the films on a scale from 1 to 5, and their EEG and physiological signals were recorded. All labels and data from each channel will be saved in separate files as part of this assignment. Rows represent channels, and columns represent timestamps for each trail. Each person's data is saved in this manner.

The tests reveal that the proposed GNN-based multimodal framework categorizes cognitive states using fMRI and EEG data. The model outperformed the conventional and DL baselines in performance metrics, revealing brain connection patterns that enable various cognitive functions. The graph-based strategy may improve neuroergonomic and therapeutic decision support by organizing and simplifying brain

connections. The system, designed for flexibility, works with many cognitive paradigms and neuroimaging datasets. This study highlights the importance of GNN topologies and multimodal fusion in cognitive neuroscience, paving the way for real-time brain monitoring systems that utilize more accurate, scalable, and interpretable models.

#### Environmental setup

Advanced high-performance computing (HPC) technology, utilizing an Ubuntu 20.04 LTS operating system, 64 GB of RAM, and an NVIDIA RTX 3090 GPU with 24 GB of video RAM, was employed for all experiments. A GNN-based multimodal fusion architecture was built using Python 3.9 and PyTorch 2.0, the most popular DL package. The numerical calculations are done using NumPy, the GNN modules with PyTorch Geometric, and the EEG and fMRI preprocessing with MNE-Python and Nilearn. Along with scikit-learn and SciPy, these libraries are utilized. Research utilized CUDA 11.8 to accelerate GPU model training. The Adam optimizer trained all models. Since the validation loss was achieved early, training was stopped to prevent overfitting. The learning rate was 0.001, and the batch size was 32. Performance indicators and fixed random seeds were calculated using 5-fold cross-validation across all libraries. To assure reproducibility and statistical reliability, this was done.

#### Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

As shown in Equation (9), the context of classification-based neural decoding systems, the accuracy rate—the percentage of correct predictions (including true positives and true negatives)—is a key performance statistic. Measuring the model's effectiveness in categorizing cognitive states across all classes is one approach to evaluating its performance. More specifically, TP shows how well the model categorizes a cognitive state as a "memory" project. Additionally, TN records occasions where the model correctly identifies the lack of a state. Acronyms for misclassifications include false positives (FP) and false negatives (FN). In imbalanced datasets, accuracy, which measures model performance, can be misleading in cognitive neuroscience because some mental states dominate the distribution. Accuracy evaluates model performance. When employed alongside accuracy, Precision, recall, and F1-score are class-sensitive measures that demonstrate the model's reliability and robustness in high-stakes domains, such as clinical diagnostics and BCI. A model's Precision assesses its performance in a specific context. Figure 3 shows accuracy.

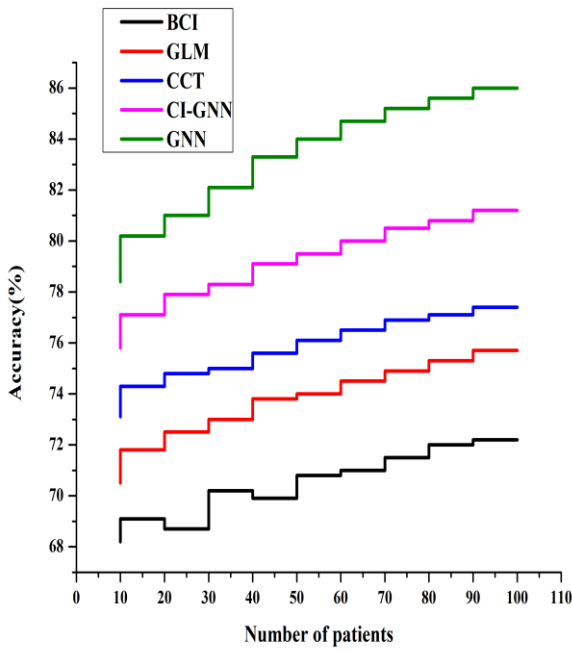


Fig. 4. Accuracy

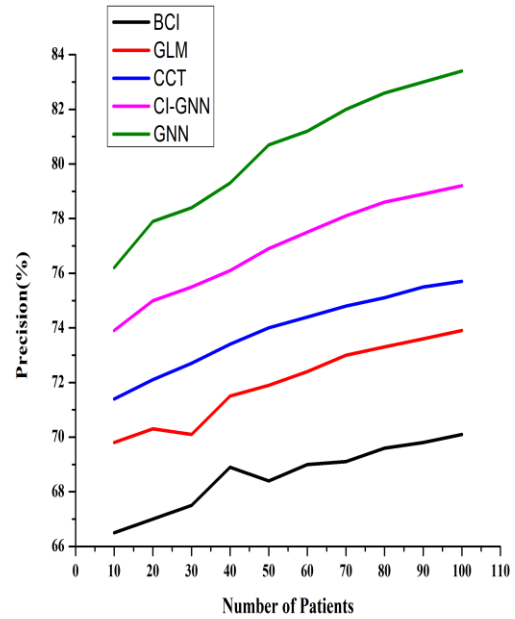


Fig. 5. Precision

**Precision**

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Equation (10) defines Precision as the model's ability to execute positive predictions. It also assesses model dependability. Using the proportion of positive occurrences (including both true and false positives) divided by the total number of positive instances, a model's prediction accuracy can be calculated. In the domain of cognitive state classification using multimodal neuroimaging data, a high precision score indicates that when the model predicts a certain cognitive state (e.g., attention or memory), it is highly likely that the prediction is correct. This becomes especially critical in real-time neurocognitive monitoring and decision-making systems, such as brain-computer interfaces (BCIs), where false positives can trigger inappropriate actions or misinterpret the user's mental intent. Therefore, Precision directly reflects the model's ability to minimize type I errors (false alarms), which is crucial for ensuring the reliability, safety, and clinical applicability of brain-state decoding architectures. However, Precision should be interpreted jointly with Recall, especially in tasks with class imbalance or where the cost of missing true cognitive events is high. Figure 4 illustrates Precision.

**Recall**

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

The Equation (11) approach, also known as sensitivity or true positive rate (TPR), measures how well a model can identify all relevant positive instances for a class. Divide the number of true positives by the sum of true positives and false negatives (FN). In the context of cognitive state classification using multimodal EEG-fMRI data, recall measures how effectively the model captures the occurrence of a specific mental state when it truly exists. High Recall indicates that the model is minimizing type II errors (i.e., missed detections), which is particularly critical in domains such as real-time brain-state monitoring, assistive neurotechnologies, or clinical diagnosis—where overlooking a cognitive event (e.g., a lapse in attention or the onset of a neurological episode) may have significant consequences. Therefore, Recall serves as a key metric for evaluating the sensitivity and responsiveness of the system, especially in scenarios involving imbalanced datasets, where some cognitive states may occur less frequently than others. For a balanced assessment of the model's reliability, Recall is typically analyzed alongside Precision and F1-score. Figure 5 shows recall

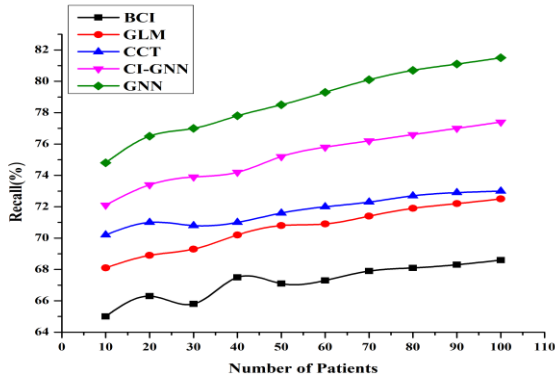


Fig. 6. Recall

**F1-Score**

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{12}$$

Equation (12) represents the F1-score, a composite evaluation metric that represents the harmonic mean of Precision and recall, providing a single measure that balances the trade-off between the two. It is particularly advantageous in imbalanced classification scenarios, which are common in cognitive state detection tasks where certain mental states may be underrepresented in the dataset. Unlike accuracy, which dominant classes can inflate, the F1-score gives a more nuanced indication of model performance by penalizing both false positives and false negatives. In this Equation, Precision reflects the correctness of positive predictions, while Recall measures the completeness of those predictions. A high F1-score implies that the model is simultaneously effective at identifying true cognitive states (high Recall) and avoiding incorrect classifications (high Precision), which is critical for reliable and interpretable brain-state decoding. This makes the F1-score an essential metric in applications such as real-time mental workload assessment, attention monitoring, and clinical neuroimaging analytics, where both detection accuracy and robustness to error are paramount. Figure 6 represents the F1-score.

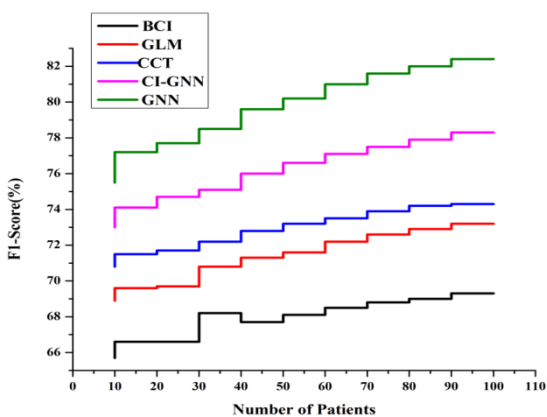


Fig. 7. F1-Score

The GNN-based multimodal framework uses classification criteria, including Accuracy, Precision, Recall, and F1-Score, to understand cognitive states from connected EEG and fMRI data. Success and toughness are possible. Through the correct categorization of several mental states, the model showed its generalization capabilities. The system's high accuracy scores indicate that it can effectively reduce false positives and provide accurate predictions of cognitive states. However, the model's high recall performance demonstrates its sensitivity to cognitive state occurrences, thereby minimizing the risk of missing detections. The model consistently generates strong F1-scores across cognitive classes, indicating that it finds a solid balance between accuracy and Recall. Because of this, it is ideal for practical applications where misclassification might cause major issues. The graph-based multimodal fusion approach is accurate, sensitive, and interpretable, making it a valuable tool for neuroergonomic and therapeutic brain-state monitoring.

GNNs are used in a structured multimodal cognitive state analysis pipeline. This pipeline uses EEG and fMRI data to improve brain-state decoding. fMRI provides spatial information from voxel-based BOLD values. Next, record brain activity using an EEG, which can identify changes over time and at high frequencies. First, we change EEG signals using spectral power analysis and entropy measurements to extract characteristics from both modalities. This lets us extract traits. Next, we use GLM/ICA or ROI-based activation to fMRI data. Using multimodal feature fusion on the feature vectors created by the method creates a single representation that aligns spatial and temporal data. Such fusion includes attention-based and tensor fusion. The next stage is to design a neural network with nodes representing EEG channels or brain regions and edges representing functional connection patterns. These patterns are commonly caused by coherence or Pearson correlation measures. The GNN receives the connection graph and unified node characteristics. After that, the network uses GCN or GAT layers to transfer and aggregate information across the graph. The network learns rich graph embeddings after training to identify cognition-related brain activity patterns. After the GNN application, the subject's attention and mental exertion are predicted. Our pipeline offers consistent and straightforward cognitive state classification through graph-based DL, facilitating the rapid integration of neuroimaging data.

**5. CONCLUSION**

A graph-based multimodal learning framework, including EEG and fMRI data, was employed in this study. To make reliable predictions of mental states, this framework was developed. The proposed model represents complex spatiotemporal brain activities by integrating the high temporal resolution of EEG with

the rich spatial information of fMRI. With GNNs, Fusion methods such as attention or tensor-based ones may enable us to extract complementary characteristics from both domains. While maintaining organic connection architectures, we may employ brain graphs. GNN-based methods surpass conventional ML and unimodal DL in accuracy, robustness, and scalability. Comparisons between the two models show this. Clinical diagnostics, neuroergonomics, and BCI may benefit from improved prediction and new insights into brain connection patterns associated with different cognitive states. Future years will have many intriguing opportunities. Adding real-time streaming EEG-fMRI data might enable online cognitive monitoring in real-world contexts. Contrastive or self-supervised learning on large amounts of unlabeled brain data may improve feature generalization. To make the architecture more time-responsive, consider spatiotemporal GNNs and graph attention networks (GATs). Physical brain atlases or individualized connectomes may improve medical interpretation of models. Ultimately, studying the model's potential application in clinical populations, such as those with neurological conditions, may lead to innovative cognitive assessment methods, early diagnosis, and targeted neurofeedback treatment approaches.

## REFERENCE

- Huang, Z., Kosan, M., Medya, S., Ranu, S., & Singh, A. (2023, February). Global Counterfactual Explainer for Graph Neural Networks. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (pp. 141-149).
- Woźniak, M., Siłka, J., & Wiczorek, M. (2023). Deep Neural Network Correlation Learning Mechanism for CT Brain Tumor Detection. *Neural Computing and Applications*, 35(20), 14611-14626.
- Niu, W., Ma, C., Sun, X., Li, M., & Gao, Z. (2023). A brain network analysis-based double-way deep neural network for emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 917-925.
- Wang, J., Li, H., Qu, G., Cecil, K. M., Dillman, J. R., Parikh, N. A., & He, L. (2023). Dynamic Weighted Hypergraph Convolutional Network for Brain Functional Connectome Analysis. *Medical image analysis*, 87, 102828.
- Wu, Q., Chen, Y., Yang, C., & Yan, J. (2023). Energy-based out-of-distribution detection for graph neural networks. arXiv preprint arXiv:2302.02914.
- Chakraborty, B., & Mukhopadhyay, S. (2023). Heterogeneous recurrent spiking neural network for spatio-temporal classification. *Frontiers in Neuroscience*, 17, 994517.
- Linka, K., Pierre, S. R. S., & Kuhl, E. (2023). Automated model discovery for the human brain using constitutive artificial neural networks. *Acta Biomaterialia*, 160, 134-151.
- Nakra, A., & Duhan, M. (2023). Deep Neural Network with Harmony Search-Based Optimal Feature Selection for EEG Signal Classification in Motor Imagery. *International Journal of Information Technology*, 15(2), 611-625.
- Chen, Z., Qing, J., & Zhou, J. H. (2023). Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36, 24841-24858.
- Tian, S., Zhu, R., Chen, Z., Wang, H., Chattun, M. R., Zhang, S., ... & Lu, Q. (2023). Prediction of suicidality in bipolar disorder using variability of intrinsic brain activity and machine learning. *Human brain mapping*, 44(7), 2767-2777.
- Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., ... & Abraham, T. (2023). Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 24705-24728.
- Lu, Y., Du, C., Zhou, Q., Wang, D., & He, H. (2023, October). Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 5899-5908).
- Singh, P., Pandey, P., Miyapuram, K., & Raman, S. (2023, June). EEG2IMAGE: image reconstruction from EEG brain signals. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
- Gao, Y., Jia, B., Houston, M., & Zhang, Y. (2023). Hybrid EEG-fNIRS brain computer interface based on common spatial pattern by using EEG-informed general linear model. *IEEE Transactions on Instrumentation and Measurement*, 72, 1-10.
- Mostafavi, A., Cruz-Garza, J. G., & Kalantari, S. (2023). Enhancing lighting design through the investigation of illuminance and correlated color temperatures' effects on brain activity: An EEG-VR approach. *Journal of Building Engineering*, 75, 106776.
- Zheng, K., Yu, S., & Chen, B. (2024). Ci-gnn: A Granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. *Neural Networks*, 172, 106147
- <https://www.kaggle.com/datasets/samnikolas/eeg-dataset>