CoRE-BOLD: Cross-Domain Robust and Equitable Ensemble for BOLD Signal Analysis

Vipul Kumar Singh[1*] VIPUL.KUMAR.SINGH@EE.IITD.AC.IN
Jyotismita Barman[1*] JYOTISMITA.BARMAN@EE.IITD.AC.IN
Sandeep Kumar[1,2,3] KSANDEEP@EE.IITD.AC.IN
Jayadeva[1,2] JAYADEVA@EE.IITD.AC.IN
[1] *Department of Electrical Engineering, Indian Institute of Technology, Delhi, India*
[2] *Yardi School of Artificial Intelligence, Indian Institute of Technology, Delhi, India*
[3] *Bharti School of Telecommunication Technology and Management, Indian Institute of Technology, Delhi, India*

Abstract

In current neuroimaging studies aimed at analysing BOLD signals, the focus has primarily been on correlation-based features derived from time series data. Considering Major Depressive Disorder (MDD), a widespread psychiatric condition, poses a complex and poorly understood pathology. Recent research has increasingly linked MDD to disruptions in brain connectivity, as observed through functional Magnetic Resonance Imaging (fMRI). Identifying the brain regions associated with neurological disorders and cognitive processes remains a central objective in neuroimaging studies. While Graph Neural Networks (GNNs) have been widely employed to extract disease-relevant information from fMRI data, existing methods face significant limitations. These limitations include neglecting the frequency-domain characteristics of neuronal interactions, inadequately incorporating non-imaging biomarkers such as sex and age, and paying insufficient attention to bias and model stability, which leaves models prone to small perturbations. We introduce CoRE-BOLD, a unified framework addressing these gaps for MDD diagnosis in this study. CoRE-BOLD employs an ensemble of stacked networks that learn complementary representations from both correlation- and coherence-based functional connectivities. To further improve the model, we enforce orthonormality constraints on the graph convolutional filters to enhance intra-network diversity and apply a diversity-maximizing regularizer for inter-network diversity. Unlike previous studies, which incorporate non-imaging sensitive attributes as biomarkers but inadvertently introduce bias, CoRE-BOLD mitigates this through a prejudice remover regularizer, promoting fairness in representation learning across both underrepresented and favored groups. Our experimental evaluation on the REST-meta-MDD dataset demonstrates the efficacy of CoRE-BOLD as a robust and fair framework for BOLD signal analysis in MDD detection, positioning it as a promising solution for real-world medical applications. Source code of CoRE-BOLD is freely available at: https://github.com/shashivipul/CoRE-BOLD.git.

Keywords: Fairness, fMRI, Functional connectivity, Graph Neural Networks, Major Depressive Disorder, Robustness, Trustworthy AI

1. Introduction

Major Depressive Disorder (MDD) is a prevalent mental health illness impacting approximately 300 million people worldwide, considerably contributing to the social and economic strains on healthcare systems (Zhou et al., 2022; Wittchen et al., 2011). The underlying neurological and pathophysiological mechanisms of MDD, like many psychiatric disorders, are largely unknown. Diagnosis often relies on subjective symptoms such as insomnia, anxiety, and emotional distress, leading to potential misdiagnosis or treatment delays (Zhang et al., 2021). Research indicates that the cortical networks in individuals with MDD show notable differences from those

* These authors contributed equally

(a) Correlation (b) Coherence

Figure 1: Top 100 significant connectivity for MDD patients in different domains.

in healthy controls, suggesting alternative diagnostic markers (Gallo et al., 2023; Kambeitz et al., 2017).

Neuronal functional connectivity, derived from the functional Magnetic Resonance Imaging (fMRI) modality, has been widely utilized to analyze the interdependencies between different brain regions (Poldrack et al., 2024). The BOLD (Blood Oxygen Level Dependent) signal measures shifts in blood oxygen levels, indicating brain activity patterns. It plays a vital role in computational neuroscience, helping to identify brain functions and assist in diagnosing neurological conditions using fMRI data. Due to the complexity of neuronal connectivity, manual analysis is difficult, prompting a shift toward automated methods to identify patterns within these networks. Traditional machine learning algorithms, like support vector machines and logistic regression, have been used to analyze features of functional connectivity(Wee et al., 2012; Khazaee et al., 2016). Recently, with the success of deep learning models in medical imaging analysis (Shen et al., 2017; Singh and Kolekar, 2022), various convolutional neural network-based models have been employed to analyze 2D representations of functional connectivity (Kawahara et al., 2017; Iqbal et al., 2023). Given the network information inherent in neuronal connectivity, Graph Neural Networks (GNNs) offer a natural and effective approach to capturing both the structural and temporal variations in functional connectivity (Li et al., 2021; Zhang et al., 2022). GNNs learn embeddings guided by node attributes, edge relationships, and the overall graph topology, effectively capturing the local and global structure of the data. This approach generalizes CNNs, extending their application from grid-like structures, such as images, to arbitrary graph-structured data.

GNN based methods have primarily been used to analyze brain functional connectivity from the temporal correlations of BOLD signals. However, temporal correlations alone are inadequate for capturing the full complexity of brain connectivity (Bastos and Schoffelen, 2016; Mohanty et al., 2020). Brain regions identified as strongly connected through temporal correlations may exhibit varying prominence under different modeling hypotheses, offering unique insights. Traditional correlation-based methods often overlook periodicity and phase relationships that frequency-domain connectivity, like coherence-based functional connectivity, can capture. This approach reveals phase relationships in the BOLD signal. Figure 1 compares the top 100 most prominent brain region connections based on both temporal correlation and coherence, demonstrating that each domain provides complementary insights into the high-dimensional brain connectome. Additionally, many studies do not sufficiently incorporate non-imaging factors like age, gender, and race, which can enhance model performance but may also introduce biases.

1.1. Related Work

GNNs have shown promise in neuroimaging for node- and graph-level tasks(Yan et al., 2019b). However, most studies have primarily focused on predictive performance metrics such as accuracy and F1-score, while neglecting crucial aspects like robustness and fairness. Transitioning deep learning from research to healthcare applications requires addressing issues of trustworthiness and vulnerability to adversarial attacks and biases among demographic groups (Zügner et al., 2018; Mujkanovic et al., 2022; Dai and Wang, 2021; Agarwal et al., 2021; Yang et al., 2024). For example, the prevalence of MDD among females highlights the need for AI systems to enhance fairness and robustness (Albert, 2015).

The medical community has increasingly focused on evaluating the trustworthiness of AI methods (Hirano et al., 2021; Ghaffari Laleh et al., 2022; Afzal et al., 2023; Chen et al., 2023; Ktena et al., 2024; Ricci Lara et al., 2022).

Recent work has examined fairness in GNNs for neurological disorders, but proposed mitigation strategies often rely on data sampling or model fine-tuning rather than addressing biases within GNN embeddings through architectural improvements(Ribeiro et al., 2022). While recent studies have

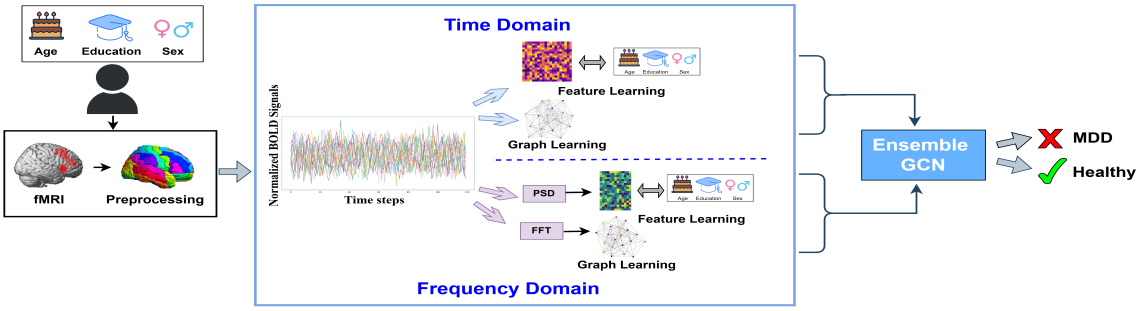


Figure 2: Overview of the proposed framework. Starting with BOLD signals, functional connectivity (FC) is computed between brain regions in both time and frequency domains. Personal information is incorporated to construct a personalized feature matrix (FM). Both the FC and FM are input into an ensemble GCN, which adaptively learns efficient embeddings to enhance prediction accuracy.

investigated fairness in neuroimaging data, they fall short in offering an algorithmic approach that simultaneously enhances both adversarial robustness and fairness (Ribeiro et al., 2022; Ghaffari Laleh et al., 2022; Hasani et al., 2022; Kaissis et al., 2020).

To the best of authors' knowledge, no previous study has integrated cross-domain functional connectivity with non-imaging attributes for brain connectome analysis. Moreover, there is a lack of tailored algorithms specifically designed to enhance both adversarial robustness and fairness in models used for analyzing brain functional connectivity.

To address these gaps, we propose a unified framework that integrates cross-domain information and demographic attributes, ensuring trustworthiness through enhanced adversarial robustness and fairness. Specifically, we employ a stacked ensemble of two GNNs with both inter-network and intra-network diversity maximization, combined with a prejudice regularizer. The key contributions of this work are summarized as follows:

① We propose CoRE-BOLD, an ensemble network for analyzing cross-domain brain functional connectivity.

② A novel strategy is introduced for jointly maximizing inter-network and intra-network diversity, leading to improved generalization and robustness.

③ A custom loss function is defined by integrating a prejudice regularizer, enhancing fairness within the framework.

④ Extensive experiments on the REST-meta-MDD dataset demonstrate the effectiveness of the proposed framework in both performance and trustworthiness.

The remainder of this paper is organized as follows: Section 2 introduces the proposed CoRE-BOLD approach. In Section 3 simulation results are presented to benchmark the proposed framework. Finally, Section 4 provides concluding remarks and discussion.

2. Proposed Method

In this section, we provide a comprehensive description of the proposed method, illustrated in Figure 2. The method consists of two main stages: (1) extracting temporal and frequency-domain functional connectivity features from the BOLD signal, and (2) training an Ensemble GNN to classify each BOLD signal.

2.1. Problem Description

For each subject S_i, the BOLD signal is represented as $\mathbf{B}_i \in \mathbb{R}^{n \times T_i}$, where n denotes the number of Region Of Interest (ROIs) and T_i is the duration of the scan. The functional connectivity matrix $\mathbf{A}_i \in \mathbb{R}^{n \times n}$ represents the relationships between ROIs in B_i. Each ROI is also associated with a d-dimensional feature vector, forming a feature matrix $\mathbf{X}_i \in \mathbb{R}^{n \times d}$. The objective is to learn a GNN model $f_{\boldsymbol{\Theta}}(\cdot)$ that maps each $\mathbf{B}_i$ to its corresponding label $\mathbf{y}_i$, where $\boldsymbol{\Theta}$ represents the model parameters.

2.2. Functional Connectivity Estimation

Estimating functional connectivity from the BOLD signal is crucial for learning graph embeddings with

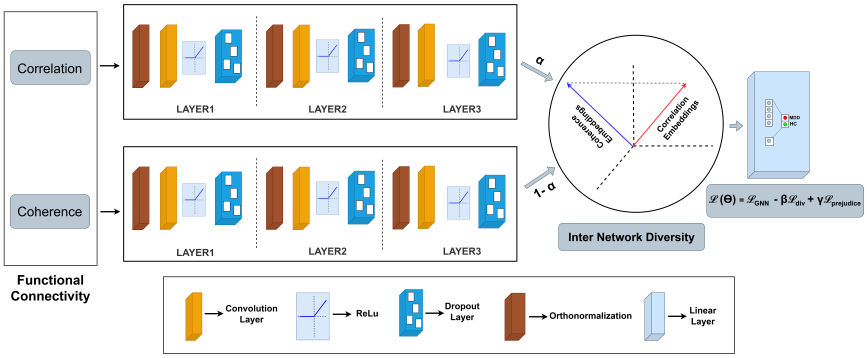


Figure 3: Architecture of the Ensemble GNN used in the CoRE-BOLD framework. Correlation and coherence graphs are input into separate networks, each incorporating a weight orthonormalization block before every graph convolutional layer to promote intra-network diversity. After the readout layers, the representations from both networks are combined using a convex combination to generate the final prediction. Inter-network diversity ensures that the networks learn complementary patterns from the two domains, enhancing the robustness and expressiveness of the model.

GNNs. For each patient S_i, we construct the ROI similarity matrix $\mathbf{C}_i \in \mathbb{R}^{n\times n}$ as detailed below:

$$\begin{bmatrix} \text{sim}(\mathbf{b}_i^{(1)}, \mathbf{b}_i^{(1)}) & \text{sim}(\mathbf{b}_i^{(1)}, \mathbf{b}_i^{(2)}) & . & \text{sim}(\mathbf{b}_i^{(1)}, \mathbf{b}_i^{(n)}) \\ . & . & . & . \\ . & . & . & . \\ \text{sim}(\mathbf{b}_i^{(n)}, \mathbf{b}_i^{(1)}) & \text{sim}(\mathbf{b}_i^{(n)}, \mathbf{b}_i^{(2)}) & . & \text{sim}(\mathbf{b}_i^{(n)}, \mathbf{b}_i^{(n)}) \end{bmatrix}$$

where, $\mathbf{b}_i^{(k)}$ is the BOLD signal for the k^{th} ROI in the i^{th} subject i.e., $\mathbf{b}_i^{(k)} = \mathbf{B}_i[k, :]$. Studies have pointed that different similarity metrics in above equation yields complementary information (Mohanty et al., 2020; Bastos and Schoffelen, 2016). The Pearson correlation coefficient is the most widely used measure to estimate similarity between regions in functional connectivity analysis. The Pearson correlation between BOLD signals from regions p and q is calculated as

$$\rho\left(\mathbf{b}_i^{(p)}, \mathbf{b}_i^{(q)}\right) = \frac{\mathbb{E}\left[\left(\mathbf{b}_i^{(p)} - \mu_i^{(p)}\right)\left(\mathbf{b}_i^{(q)} - \mu_i^{(q)}\right)\right]}{\sigma_{\mathbf{b}_i^{(p)}}\sigma_{\mathbf{b}_i^{(q)}}} \quad (1)$$

where, $\mu_i^{(p)}$ and $\sigma_{\mathbf{b}_i^{(p)}}$ is the mean and standard deviation of the p^{th} BOLD signal. In this work, we further employ partial correlation, derived from the Pearson correlation defined earlier, to quantify the similarity between pairs of ROIs.

Once the pairwise correlations are computed, a non-negative threshold is applied to generate the final brain functional connectivity matrix $\mathbf{A}_i$ for each subject.

Functional connectivity based on temporal correlations often overlooks patterns associated with the rhythmic oscillations of neurons. To better capture periodic interactions in the brain, it is necessary to transform the BOLD signal into the frequency domain and analyze the resulting similarities. In this study, we compute frequency-domain functional connectivity using coherence statistics, defined between the BOLD signals from regions p and q as

$$\text{Coh}_{\mathbf{b}_i^{(p)},\mathbf{b}_i^{(q)}}(f) = \frac{\left|\text{C}_{\mathbf{b}_i^{(p)},\mathbf{b}_i^{(q)}}(f)\right|^2}{\text{C}_{\mathbf{b}_i^{(p)},\mathbf{b}_i^{(p)}}(f)\ \text{C}_{\mathbf{b}_i^{(q)},\mathbf{b}_i^{(q)}}(f)} \quad (2)$$

where, numerator represents the cross-spectral density between the signals of the ROIs, while the denominator is the product of their individual auto-spectral densities. We propose a joint approach that leverages both cross-domain functional connectivity to comprehensively analyze the BOLD signals, capturing complementary insights from both temporal and frequency domains.

2.3. Diversity Maximizing and Prejudice Minimizing Ensemble GNN

This study presents a novel approach for cross-domain functional connectivity analysis, employing a stacking ensemble strategy to improve the understanding of topological patterns within brain net-

works. By integrating two models, this method leverages their distinct capabilities in processing and analyzing cross-domain data. The approach facilitates a deeper and more nuanced understanding of brain connectomes through diverse feature extraction and fusion techniques. Ensemble models provide an effective solution in handling out of distribution robustness (Ovadia et al., 2019). The core stacking ensemble framework is illustrated in Figure 3.

Conventional GNNs typically employ identical graph filters across layers, often resulting in graph representations that are either shifted versions of one another or nearly indistinguishable. This uniformity reduces diversity in the feature space and limits the richness of the learned representations (Ayinde et al., 2019; Choudhary et al., 2023). As a result, the network's expressivity and capacity are constrained, weakening its ability to capture complex patterns and compromising its robustness. Furthermore, inter-network diversity is crucial for effective ensembling, as it ensures the capture of complementary information from each domain.

2.3.1. Intra and Inter-Network Diversity

Orthogonal filters have been widely studied in signal processing due to their ability to preserve activation energy and minimize redundancy in representations (Zhou et al., 2006; Wang et al., 2020; Huang et al., 2018). To enhance intra-network diversity, we enforce the learning of orthonormal graph filters within each layer of both networks by constraining the product of the parameter matrix $\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}$ to be equal to the identity matrix.

Given M ensemble networks with parameters $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_{\mathrm{M}}$, the output representations from the t^{th} network for an input graph-feature pair $(\mathbf{A}_{\mathrm{i}}, \mathbf{X}_{\mathrm{i}})$ are $\mathrm{f}_{\boldsymbol{\Theta}_{\mathrm{t}}}(\mathbf{A}_{\mathrm{i}}, \mathbf{X}_{\mathrm{i}})$. The final prediction of the ensemble is determined by the convex combination of embeddings from each network. To quantify the diversity between the two networks, we define the weighted diversity as:

$$\mathcal{D}(\boldsymbol{\Theta}_{\mathrm{s}}, \boldsymbol{\Theta}_{\mathrm{t}}) = \delta\left(\alpha_{\mathrm{i}} \mathrm{f}_{\boldsymbol{\Theta}_{\mathrm{s}}}(\mathbf{A}_{\mathrm{i}}, \mathbf{X}_{\mathrm{i}}), \alpha_{\mathrm{j}} \mathrm{f}_{\boldsymbol{\Theta}_{\mathrm{t}}}(\mathbf{A}_{\mathrm{i}}, \mathbf{X}_{\mathrm{i}})\right). \quad (3)$$

We consider $\delta(\cdot)$ as the cosine-similarity metric defined as

$$\delta(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathrm{x}\| \, \|\mathrm{y}\|}. \quad (4)$$

While we introduce weighted diversity, note that when using cosine similarity, the weights assigned to each model do not influence the result due to the normalization step. However, this framework is more general and can accommodate other distance metrics, where the weights would play a more significant role. Consequently, we define the inter-network diversity loss as:

$$\mathcal{L}_{\mathrm{div}} = \sum_{\mathrm{t}=1}^{\mathrm{M}} \sum_{\mathrm{s} \neq \mathrm{t}} \mathcal{D}(\boldsymbol{\Theta}_{\mathrm{s}}, \boldsymbol{\Theta}_{\mathrm{t}}). \quad (5)$$

In the Ensemble GNN depicted in Figure 3, we consider an ensemble of two networks, hence $\mathrm{M} = 2$. To promote diversity within the ensemble network, the objective defined in (5) must be maximized.

2.3.2. Prejudice Regularizer

In critical AI systems, user data often contains sensitive demographic attributes such as sex or race, which may lead to unintended biases in the trained model's predictions. For a sensitive attribute S, prejudice refers to the unwanted statistical dependence between S and the model's outputs. To mitigate bias towards any specific group based on S, it is crucial to ensure consistency in model performance across all groups. In the case of a binary sensitive attribute defining two groups G_0 and G_1 the discrepancy in performance between the groups can be quantified by the following unfairness loss function:

$$\mathcal{L}_{\mathrm{prejudice}} = \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathrm{G}_0} \left[\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right)\right] - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathrm{G}_1} \left[\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right)\right] \right| \quad (6)$$

where, $\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right)$ represents the prediction loss function (e.g., cross-entropy), and the expectation is taken over the samples belonging to each group. To handle cases with multiple binary sensitive attributes, the loss function can be generalized by summing individual prejudice losses. Specifically, if there are u binary sensitive attributes, the loss can be formulated as:

$$\mathcal{L}_{\mathrm{prejudice}} = \sum_{\mathrm{i}=1}^{\mathrm{u}} \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathrm{G}_{0_{\mathrm{i}}}} \left[\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right)\right] - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathrm{G}_{1_{\mathrm{i}}}} \left[\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right)\right] \right| \quad (7)$$

where, $\mathrm{G}_{0_{\mathrm{i}}}$ and $\mathrm{G}_{1_{\mathrm{i}}}$ represent the two groups corresponding to the i^{th} sensitive attribute.

2.3.3. Learning Algorithm

We consider a training set consisting of graph-feature pairs with associated labels $\{(\mathbf{A}_{\mathrm{i}}, \mathbf{X}_{\mathrm{i}}), \mathbf{y}_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{L}}$. To learn the GNN hypothesis function $\mathrm{f}_{\boldsymbol{\Theta}}$, the data is processed through multiple layers of graph filters.

Each standard GNN layer consists of two primary operations: message passing between graph nodes and feature aggregation. These operations are mathematically summarized as:

$$\mathbf{e}_{\mathrm{u}}^{(\mathrm{k})} = \sigma\left(\boldsymbol{\Theta}_{\mathrm{self}}^{(\mathrm{k})}\mathbf{e}_{\mathrm{u}}^{(\mathrm{k}-1)} + \boldsymbol{\Theta}_{\mathrm{neighbor}}^{(\mathrm{k})}\sum_{\mathrm{v}\in\mathcal{N}_{\mathrm{u}}}\mathbf{e}_{\mathrm{v}}^{(\mathrm{k}-1)} + \mathbf{b}^{(\mathrm{k})}\right) \tag{8}$$

where $\mathbf{e}_{\mathrm{u}}^{(\mathrm{k})}$ denotes the embedding of node u after k rounds of GNN operations, and $\mathcal{N}_{\mathrm{u}}$ represents the neighborhood of node u. The parameters $\boldsymbol{\Theta}_{\mathrm{self}}^{(\mathrm{k})}$, $\boldsymbol{\Theta}_{\mathrm{neighbor}}^{(\mathrm{k})}$, and $\mathbf{b}^{(\mathrm{k})}$ are learnable. The GNN hypothesis function is represented as $\mathrm{f}_{\boldsymbol{\Theta}}(\cdot)$. Under this notion the learnable parameter $\boldsymbol{\Theta}$ consists parameters of each layer i.e., $\boldsymbol{\Theta} = \left\{\boldsymbol{\Theta}^{(\mathrm{k})}, \mathbf{b}^{(\mathrm{k})} \;;\; \mathrm{k} = 1, 2, \dots\right\}$. The network is trained to minimize the GNN loss function $\mathcal{L}_{\mathrm{GNN}}\left(\mathbf{y}, \mathrm{f}_{\boldsymbol{\Theta}}\left(\mathbf{A}, \mathbf{X}\right)\right)$.

In this work, we propose a method for learning a robust and fair GNN specifically designed for cross-domain functional connectivity analysis of the brain. To achieve this, we solve the following constrained optimization problem:

$$\begin{aligned}\boldsymbol{\Theta}^{\star} = \arg\min_{\boldsymbol{\Theta}} \;& \mathcal{L}_{\mathrm{GNN}} - \beta\mathcal{L}_{\mathrm{div}} + \gamma\mathcal{L}_{\mathrm{prejudice}} \\ \text{s.t.,}\; & \boldsymbol{\Theta}^{(\mathrm{k})} \in \mathcal{S}_{\boldsymbol{\Theta}} \;\; \forall\, \mathrm{k} = 1, 2, \dots\end{aligned} \tag{9}$$

where, $\mathcal{S}_{\boldsymbol{\Theta}}$ is set consisting of matrix family over Stiefel manifold. We address the constraints of problem (9) by explicitly orthonormalizing the graph filters at each layer during every forward pass, utilizing the fast iterative algorithm introduced by Björck and Bowie (1971). This method iteratively computes an orthonormal matrix for a given input matrix $\mathbf{W}$ by applying the following update:

$$\mathbf{W}_{\mathrm{t}+1} = \mathbf{W}_{\mathrm{t}}\left(\mathbf{I} + \frac{1}{2}\mathbf{Q}_{\mathrm{t}} + \cdots + (-1)^{\mathrm{p}}\binom{-1/2}{\mathrm{p}}\mathbf{Q}_{\mathrm{t}}^{\mathrm{p}}\right), \tag{10}$$

where $\mathbf{Q}_{\mathrm{t}} = \mathbf{I} - \mathbf{W}_{\mathrm{t}}^{\mathrm{T}}\mathbf{W}_{\mathrm{t}}$ and p denotes the order of the Taylor series expansion. The convergence rate improves as p increases. In our experiments, we set $\mathrm{p} = 1$ and perform 15 iterations. Hence, the unconstrained problem of (9) can be effectively minimized using stochastic gradient descent since, all the objective functions are differentiable. The gradient of total loss w.r.t. parameters will be

$$\nabla_{\boldsymbol{\Theta}}\mathcal{L}\left(\boldsymbol{\Theta}\right) = \frac{\partial}{\partial\boldsymbol{\Theta}}\mathcal{L}_{\mathrm{GNN}} - \beta\frac{\partial}{\partial\boldsymbol{\Theta}}\mathcal{L}_{\mathrm{div}} + \gamma\frac{\partial}{\partial\boldsymbol{\Theta}}\mathcal{L}_{\mathrm{prejudice}} \tag{11}$$

The stochastic gradient step for the update of parameters will be

$$\boldsymbol{\Theta}_{\mathrm{t}+1} = \boldsymbol{\Theta}_{\mathrm{t}} - \eta\left(\nabla_{\boldsymbol{\Theta}}\mathcal{L}\left(\boldsymbol{\Theta}_{\mathrm{t}}\right)\right). \tag{12}$$

The detailed pseudo-code of the learning algorithm is given in the Algorithm 1.

Algorithm 1: CoRE-BOLD Algorithm

Input: Training data $\{(\mathbf{A}_{\mathrm{i}}, \mathbf{X}_{\mathrm{i}}), \mathbf{y}_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{L}}$
Parameters: β, γ, η
Output: Robust and fair model $\boldsymbol{\Theta}$
▶ Set $\mathrm{t} = 0$ and initialize $\boldsymbol{\Theta}_0$
while *stopping criteria not met* **do**
  $\boldsymbol{\Theta}_{\mathrm{t}} \leftarrow$ **Björck-Orthonormalization**() using (10)
  $\mathcal{L}_{\mathrm{GNN}}, \mathcal{L}_{\mathrm{div}}, \mathcal{L}_{\mathrm{prejudice}} \leftarrow$ **Forward**()
  $\nabla_{\boldsymbol{\Theta}}\mathcal{L}\left(\boldsymbol{\Theta}_{\mathrm{t}}\right) \leftarrow$ **Gradient**() using (11)
  $\boldsymbol{\Theta}_{\mathrm{t}+1} \leftarrow$ **Update**() using (12)
end

3. Experimental Results

In this section, we present numerical simulations conducted on the REST-meta-MDD dataset[1] (Yan et al., 2019a). First, we evaluate the predictive performance of CoRE-BOLD in comparison to baseline models. Next, we analyze the robustness and fairness of CoRE-BOLD through a series of detailed experiments. All the analysis is done with 5 fold cross validation. The backbone GNN model for CoRE-BOLD is Graph Convolutional Network (GCN). The base code was implemented using PyTorch Geometric, a deep learning library built on the PyTorch framework, with Python 3.9.18 and PyTorch 2.1.2.

3.0.1. Dataset Description

The dataset consists of resting-state fMRI scans from 1300 MDD patients and 1128 Healthy Controls (HCs) collected from 17 hospitals across China (Yan et al., 2019a). Each site contributed, on average, 52.0±52.4 MDD patients (range 13-282) and 45.1±46.9 HCs (range 6-251). We have considered 1570 data samples from all the sites following (Yan et al., 2019a).

1. https://rfmri.org/REST-meta-MDD

Table 1: Performance on prediction task: Accuracy (Acc), Area Under Curve (AUC) and F1-score with mean ± standard deviation (%). The best results are in bold.

	Atlas: AAL			Atlas: Dosenbach		
Method	**Acc**	**AUC**	**F1**	**Acc**	**AUC**	**F1**
SVM	54.80 ± 2.76	54.30 ± 3.67	58.60 ± 2.50	55.40 ± 3.33	55.20 ± 4.01	63.70 ± 3.19
LR	53.90 ± 3.59	54.50 ± 3.65	57.10 ± 4.03	54.10 ± 3.63	53.50 ± 3.57	60.50 ± 3.63
RF	52.20 ± 4.10	54.70 ± 4.25	53.90 ± 5.15	54.70 ± 4.40	55.20 ± 3.78	59.50 ± 2.83
GAT	57.90 ± 9.97	62.13 ± 3.07	60.22 ± 3.53	58.00 ± 3.89	61.54 ± 5.68	58.74 ± 6.80
GIN	60.64 ± 2.97	63.92 ± 2.53	61.66 ± 2.53	53.44 ± 3.32	55.82 ± 4.81	54.57 ± 3.99
GCN(Correlation)	55.22 ± 4.24	57.58 ± 4.88	60.79 ± 8.20	53.09 ± 3.23	56.27 ± 4.13	52.86 ± 7.58
GCN(Coherence)	54.39 ± 3.44	56.07 ± 3.73	61.74 ± 4.65	52.84 ± 2.16	54.06 ± 4.14	52.01 ± 2.16
BrainNetCNN	59.62 ± 2.84	59.57 ± 2.88	59.36 ± 2.85	56.56 ± 3.78	56.54 ± 3.66	55.69 ± 4.08
TGCN	57.35 ± 2.06	62.65 ± 3.02	67.28 ± 1.81	55.76 ± 0.98	61.61 ± 2.66	47.51 ± 1.00
BrainGNN	53.38 ± 2.73	53.18 ± 2.87	56.69 ± 3.12	54.85 ± 3.53	54.41 ± 3.52	60.16 ± 4.68
Weighted Ensemble	63.12 ± 4.20	$\mathbf{68.08 \pm 4.30}$	65.65 ± 6.13	62.67 ± 2.39	$\mathbf{67.74 \pm 3.15}$	65.21 ± 5.08
CoRE-BOLD	$\mathbf{65.08 \pm 3.28}$	67.88 ± 4.39	$\mathbf{69.41 \pm 2.48}$	$\mathbf{64.93 \pm 1.54}$	66.17 ± 1.00	$\mathbf{68.55 \pm 2.78}$

Detailed statistical overview of the considered dataset is provided in Appendix A, while the pre-processing pipeline is detailed in Appendix A.1.

To evaluate fairness, we grouped the data based on key demographic attributes. For age, we created two groups: 18–30 and 31–60 years. For education, we divided participants into those with fewer than 14 years of total education and those with more than 14 years.

3.0.2. Baselines

For baseline comparisons, we considered several machine learning models, including SVM, Random Forest, and Logistic Regression, using vectorized correlation matrices as input features. Additionally, we evaluated a range of deep learning architectures such as GCN (Kipf and Welling, 2016), GAT (Velickovic et al., 2017), GIN (Xu et al., 2018), TGCN (Dai et al., 2023), BrainGNN (Li et al., 2021), and BrainNetCNN (Kawahara et al., 2017), all of which utilize correlation-based functional connectivity as input, following their original implementations. For comparison we show results of GCN on using coherence-based connectivity matrix. We also included a variant of the proposed CoRE-BOLD method, where layer-wise orthonormalization is omitted, and both regularization terms β and γ are set to zero. We refer to this variant as Weighted Ensemble. Comparing with the Weighted Ensemble will assess the impact of our approach on maximizing both intra- and inter-network diversity.

3.1. Results

3.1.1. Predictive Performance

Quantitative results on the predictive performance of all baseline methods and the proposed framework are summarized in the Table 1. We evaluated performance using two brain atlases: AAL and Dosenbach. While the baselines perform less favorably with the Dosenbach atlas, our proposed framework achieves comparable performance across both atlases. Deep learning methods generally outperform traditional machine learning methods, highlighting the value of advanced representation learning with complex neural networks. BrainNetCNN uses heatmap corresponding to correlation matrix. Among state-of-the-art methods with a GCN backbone, TGCN exhibits the highest predictive performance. It can be noted that GCN network trained on coherence functional connectivity achieves similar performance to that of GCN trained on correlation matrix. Our proposed Weighted Ensemble method significantly surpasses all other methods. Further enhancements are observed when incorporating inter-network and intra-network diversity and prejudice regularization within the CoRE-BOLD framework. These results demonstrate that leveraging complementary informa-

Table 2: Performance Comparison of Proposed Approach with Different Prejudice Regularization

Models	**AUC-ROC (↑)**	**SER (↓)**		
		Sex	**Education**	**Age**
GCN	57.58 ± 4.88	1.16 ± 0.28	1.10 ± 0.10	1.11 ± 0.30
GIN	63.92 ± 2.53	2.04 ± 1.27	1.33 ± 0.19	1.33 ± 0.19
GAT	62.13 ± 3.07	1.71 ± 0.53	1.67 ± 1.46	1.52 ± 0.48
BrainGNN	53.18 ± 2.87	1.36 ± 0.25	1.24 ± 0.33	1.44 ± 0.43
BrainNetCNN	59.57 ± 2.88	3.14 ± 1.85	3.47 ± 4.12	5.63 ± 4.45
TGCN	62.65 ± 3.02	1.27 ± 0.09	1.10 ± 0.05	1.10 ± 0.05
Weighted Ensemble	68.08 ± 4.30	1.19 ± 0.19	1.43 ± 0.27	1.23 ± 0.14
CoRE-BOLD (Sex)	67.88 ± 4.39	1.003 ± 0.002	1.008 ± 0.003	1.004 ± 0.004
CoRE-BOLD (Age)	67.09 ± 0.02	1.003 ± 0.002	1.0068 ± 0.006	1.005 ± 0.003
CoRE-BOLD (Education)	$\mathbf{68.72 \pm 4.22}$	1.003 ± 0.002	1.009 ± 0.007	1.005 ± 0.004
CoRE-BOLD (None)	66.10 ± 4.76	1.242 ± 0.183	1.341 ± 0.212	1.472 ± 0.238
CoRE-BOLD (All)	66.57 ± 5.56	$\mathbf{1.001 \pm 0.001}$	$\mathbf{1.002 \pm 0.001}$	$\mathbf{1.002 \pm 0.002}$

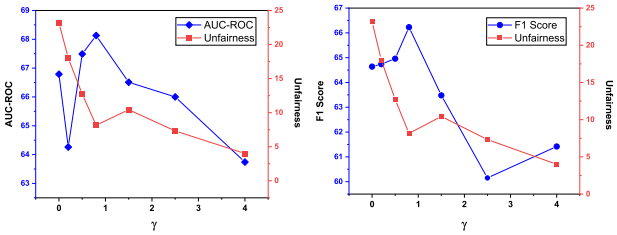
tion from multiple domains, along with demographic data, leads to more effective representation learning. Analysis in further subsections is restricted on the AAL atlas.

3.1.2. Fairness Evaluation

The results in the previous section highlight the significance of incorporating demographic information for predictive performance. However, the use of sensitive demographic data can introduce biases against disparate groups. Additionally, algorithmic biases may emerge due to the dataset's imbalance, with a higher prevalence of female MDD samples. In this section, we conduct a comprehensive investigation into the fairness of the proposed framework for MDD prediction. We adopt Skewed Error Rate (SER) to benchmark the fairness score (Wang and Deng, 2020).

$$\text{SER} = \frac{\max\limits_{g} \quad P_g}{\min\limits_{g} \quad P_g} \tag{13}$$

where, P_g is the performance of algorithm on g^{th} group. Table 2 compares the SER results of all methods. Additionally, for counterfactual fairness evaluation we have used Unfairness value metric, defined as the ratio of samples for which the model's prediction changes when the sensitive attribute is altered. The counterfactual fairness analysis results are presented in Figure 4, showing a clear trade-off between predictive performance and Unfairness. As the regularization parameter γ increases, penalizing the model for disparate performance across groups, Unfairness is significantly reduced. This framework provides medical practitioners with precise control over the balance between predictive performance and fairness, allowing them to adjust the model based on the desired emphasis.



(a) AUC-ROC (b) F1-score

Figure 4: Analysis of Trade-off between predictive performance and fairness of proposed framework.

3.1.3. Robustness Evaluation

In analyzing the trustworthiness of the proposed framework, we evaluate institutional generalization using the Leave-One-Site-Out (LOSO) cross-validation setting. Given that our dataset is collected from 16 sites, we trained the model on data from 15 sites while testing on the 16th. This process was repeated so that each site was used once for testing. The average results of all the evaluations is presented in Figure 5. The results indicate that our model demonstrates robustness for institutional gen-

Table 3: Ablation study of different modules in CoRE-BOLD.

Methods	Accuracy (↑)	F1 Score (↑)	AUC-ROC (↑)	Unfairness (↓)
CoRE-BOLD	**65.08 ± 3.28**	**69.41 ± 2.48**	**67.88 ± 4.39**	08.15 ± 3.78
Remove Intra-network diversity	61.02 ± 2.37	63.48 ± 4.27	66.51 ± 3.28	10.44 ± 7.83
Remove Inter-network diversity	61.02 ± 3.85	63.03 ± 2.95	66.55 ± 3.97	**04.33 ± 3.65**
Remove Prejudice regularizer	62.17 ± 3.03	64.64 ± 3.36	66.79 ± 4.71	23.18 ± 5.13

Table 4: Ablation Study of Using Demographics for Classification

Models	Accuracy (↑)	F1 Score (↑)	AUC-ROC (↑)	SER (↓)
GCN w/o demographics (Time)	55.22 ± 4.24	60.79 ± 8.20	57.58 ± 4.88	1.16 ± 0.28
GCN w/ demographics (Time)	61.29 ± 4.15	66.67 ± 3.42	65.72 ± 5.14	1.52 ± 0.56
GCN w/o demographics (Frequency)	54.39 ± 3.44	56.07 ± 3.73	61.74 ± 4.65	1.13 ± 0.11
GCN w/ demographics (Frequency)	61.42 ± 4.49	65.90 ± 3.14	67.39 ± 5.32	1.34 ± 0.44
Weighted Ensemble w/o demographics	56.48 ± 2.81	63.31 ± 6.49	53.20 ± 3.80	1.17 ± 0.18
Weighted Ensemble w/ demographics	63.12 ± 4.25	65.65 ± 6.13	68.08 ± 4.30	1.19 ± 0.19
CoRE-BOLD w/o demographics	57.21 ± 2.84	63.65 ± 4.17	52.45 ± 3.43	1.18 ± 0.14
CoRE-BOLD	**65.08 ± 3.28**	**69.41 ± 2.48**	**67.88 ± 4.39**	**1.003 ± 0.002**

eralization, highlighting its effectiveness in practical scenarios. Additional results by adding noise in the BOLD signals and demographic attributes are presented in the Appendix C.

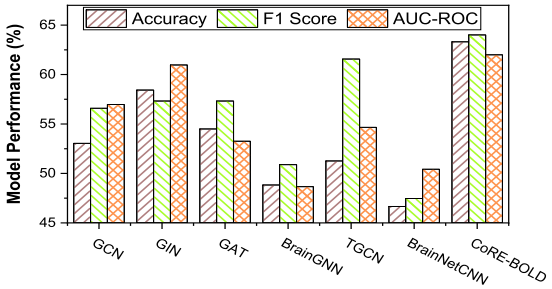


Figure 5: Comparative analysis of institute generalization

3.1.4. Ablation Study

We conducted ablation studies to assess the contribution of key modules within the CoRE-BOLD framework and to better understand the factors behind its superior performance. The results in the Table 3 compare the full CoRE-BOLD model with versions where individual components are removed. These results underscore the importance of each module in achieving optimal performance. Additionally, we performed a ablation study of the importance of using demographic information. The results are presented in Table 4. Sensitivity analysis of hyperparameter is performed in the Appendix E.

4. Conclusion

In this paper, we introduce CoRE-BOLD, a unified framework for robust and fair analysis of brain connectomes. CoRE-BOLD leverages cross-domain functional connectivity to train an ensemble network that maximizes diversity and minimizes prejudice. Our experiments on MDD classification demonstrate its superior performance compared to baselines. Additionally, we show that the framework is resilient to adversarial noise and effectively reduces algorithmic biases arising from sensitive demographic information. CoRE-BOLD lays a solid foundation for advancing trustworthy AI in the analysis of BOLD signals.

Acknowledgment

Vipul Kumar Singh is supported by Prime Minister's Research Fellowship (1402107), Jyotismita Barman is supported by Tata Consultancy Services Research Fellowship, and Sandeep Kumar is supported by Tower ML for Social Good Health (RP04636N).

References

Muhammad Muneeb Afzal, Muhammad Osama Khan, and Shujaat Mirza. Towards equitable kidney tumor segmentation: Bias evaluation and mitigation. In *Machine Learning for Health (ML4H)*, pages 13–26. PMLR, 2023.

Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pages 2114–2124. PMLR, 2021.

Paul R Albert. Why is depression more prevalent in women?, 2015.

Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada. Regularizing deep neural networks by enhancing diversity in feature extraction. *IEEE transactions on neural networks and learning systems*, 30(9):2650–2661, 2019.

André M Bastos and Jan-Mathijs Schoffelen. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, 9:175, 2016.

Åke Björck and Clazett Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8 (2):358–364, 1971.

Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

Anshul Choudhary, Anil Radhakrishnan, John F Lindner, Sudeshna Sinha, and William L Ditto. Neuronal diversity can improve machine learning for physics and beyond. *Scientific Reports*, 13(1): 13962, 2023.

Lulu Cui, Shu Li, Siman Wang, Xiafang Wu, Yingyu Liu, Weiyang Yu, Yijun Wang, Yong Tang, Maosheng Xia, and Baoman Li. Major depressive disorder: hypothesis, mechanism, prevention and treatment. *Signal Transduction and Targeted Therapy*, 9(1):30, 2024.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 680–688, 2021.

Peishan Dai, Da Lu, Yun Shi, Ying Zhou, Tong Xiong, Xiaoyan Zhou, Zailiang Chen, Beiji Zou, Hui Tang, Zhongchao Huang, et al. Classification of recurrent major depressive disorder using a new time series feature extraction method through multisite rs-fmri data. *Journal of Affective Disorders*, 339:511–519, 2023.

Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

Selene Gallo, Ahmed El-Gazzar, Paul Zhutovsky, Rajat M Thomas, Nooshin Javaheripour, Meng Li, Lucie Bartova, Deepti Bathula, Udo Dannlowski, Christopher Davey, et al. Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies. *Molecular Psychiatry*, 28(7):3013–3022, 2023.

Narmin Ghaffari Laleh, Daniel Truhn, Gregory Patrick Veldhuizen, Tianyu Han, Marko van Treeck, Roman D Buelow, Rupert Langer, Bastian Dislich, Peter Boor, Volkmar Schulz, et al. Adversarial attacks and adversarial robustness in computational pathology. *Nature communications*, 13(1):5711, 2022.

Navid Hasani, Michael A Morris, Arman Rahmim, Ronald M Summers, Elizabeth Jones, Eliot Siegel, and Babak Saboury. Trustworthy artificial intelligence in medical imaging. *PET clinics*, 17(1):1–12, 2022.

Hokuto Hirano, Akinori Minagi, and Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21:1–13, 2021.

Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Saeed Iqbal, Adnan N. Qureshi, Jianqiang Li, and Tariq Mahmood. On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Archives of Computational Methods in Engineering*, 30(5): 3173–3233, 2023.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311, 2020.

Joseph Kambeitz, Carlos Cabral, Matthew D Sacchet, Ian H Gotlib, Roland Zahn, Mauricio H Serpa, Martin Walter, Peter Falkai, and Nikolaos Koutsouleris. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. *Biological psychiatry*, 82 (5):330–338, 2017.

Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

Ali Khazaee, Ata Ebrahimzadeh, and Abbas Babajani-Feremi. Application of advanced machine learning methods on resting-state fmri network for identification of mild cognitive impairment and alzheimer's disease. *Brain imaging and behavior*, 10:799–817, 2016.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, pages 1–8, 2024.

Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

Rosaleena Mohanty, William A Sethares, Veena A Nair, and Vivek Prabhakaran. Rethinking measures of functional connectivity via feature extraction. *Scientific reports*, 10(1):1298, 2020.

Felix Mujkanovic, Simon Geisler, Stephan Günnemann, and Aleksandar Bojchevski. Are defenses for graph neural networks robust? *Advances in Neural Information Processing Systems*, 35:8954–8968, 2022.

C Otte, S Gold, B Penninx, C Pariante, A Etkin, M Fava, DC Mohr, and AF Schatzberg. Major depressive disorder. nature reviews disease primers. 2016; 2: 16065.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2024.

Fernanda Ribeiro, Valentina Shumovskaia, Thomas Davies, and Ira Ktena. How fair is your graph? exploring fairness concerns in neuroimaging studies. In *Machine Learning for Healthcare Conference*, pages 459–478. PMLR, 2022.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *nature communications*, 13(1):4581, 2022.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

Vipul Kumar Singh and Maheshkumar H Kolekar. Deep learning empowered covid-19 diagnosis using chest ct scan images for collaborative edge-cloud computing platform. *Multimedia Tools and Applications*, 81(1):3–30, 2022.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations

in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11505–11515, 2020.

Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

Chong-Yaw Wee, Pew-Thian Yap, Daoqiang Zhang, Kevin Denny, Jeffrey N Browndyke, Guy G Potter, Kathleen A Welsh-Bohmer, Lihong Wang, and Dinggang Shen. Identification of mci individuals using structural and functional connectivity networks. *Neuroimage*, 59(3):2045–2056, 2012.

Hans-Ulrich Wittchen, Frank Jacobi, Jürgen Rehm, Anders Gustavsson, Mikael Svensson, Bengt Jönsson, Jes Olesen, Christer Allgulander, Jordi Alonso, Carlo Faravelli, et al. The size and burden of mental disorders and other disorders of the brain in europe 2010. *European neuropsychopharmacology*, 21(9):655–679, 2011.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019a.

Chaogan Yan and Yufeng Zang. Dparsf: a matlab toolbox for" pipeline" data analysis of resting-state fmri. *Frontiers in systems neuroscience*, 4:1377, 2010.

Yujun Yan, Jiong Zhu, Marlena Duda, Eric Solarz, Chandra Sripada, and Danai Koutra. Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 772–782, 2019b.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, pages 1–11, 2024.

Hao Zhang, Ran Song, Liping Wang, Lin Zhang, Dawei Wang, Cong Wang, and Wei Zhang. Classification of brain disorders in rs-fmri via local-to-global graph neural networks. *IEEE transactions on medical imaging*, 42(2):444–455, 2022.

Yu Zhang, Wei Wu, Russell T Toll, Sharon Naparstek, Adi Maron-Katz, Mallissa Watts, Joseph Gordon, Jisoo Jeong, Laura Astolfi, Emmanuel Shpigel, et al. Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography. *Nature biomedical engineering*, 5(4):309–323, 2021.

Jianping Zhou, Minh N Do, and Jelena Kovacevic. Special paraunitary matrices, cayley transform, and multidimensional orthogonal filter banks. *IEEE Transactions on Image Processing*, 15(2): 511–519, 2006.

Zhiyuan Zhou, Yanrong Guo, Shijie Hao, and Richang Hong. Hierarchical multifeature fusion via audio-response-level modeling for depression detection. *IEEE transactions on computational social systems*, 10(5):2797–2805, 2022.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2847–2856, 2018.

Appendix A. Dataset Description

The dataset utilized in this study originates from the Rest-meta-MDD project and includes demographic attributes such as gender, age, and education levels. Data selected from 16 distinct sites, with detailed descriptions and statistical summaries provided in Table A.1.

A.1. Data Pre-processing

To eliminate variability, the T1 weighted rs -fMRI images were pre-processed using a standardized pre-processing pipeline Data Processing Assistant for Resting-State fMRI (DPARSF) (Yan and Zang, 2010) at each local participating site. The pre-processing steps are **(i) Slice Timing Correction:** To synchronize all slices to a reference time-point, **(ii)Realignment:** To correct the functional data for a movement that has occurred during scanning, **(iii) Removing Covariates:** here we used data with global signal regression, **(iv) Spatial Normalization:** it is done to ensure consistent anatomical coordinates across all the subjects, **(v) Co registration:** to put the structural and functional data in alignment, **(vi) band-pass filtering:** it is done with (0.01 - 0.1Hz), **(vii) Normalization:** it is done normalized to a symmetric template and finally smoothened to enhance the signal and cancel out any noise. Following pre-processing, time series BOLD signals corresponding to specific brain atlases are extracted. Subsequently, poor-quality and incomplete BOLD signals are also excluded to maintain data integrity across two atlases: Anatomical Automatic Labeling (AAL - 116 ROIs) (Tzourio-Mazoyer et al., 2002), Dosenbach's 160 functional ROIs(Dosenbach et al., 2010).

Appendix B. Complexity Analysis

The worst case computational complexity of CoRE-BOLD algorithm primarily is dominated by the orthonormalization process used in the forward pass and the training of the ensemble GNN using the proposed loss function, which includes both diversity maximization and prejudice minimization terms. In our experiments, we observed that applying first-order orthonormalization to the layer-wise weights offers a good balance between performance and complexity. Considering a GNN with L layers, an input graph with n nodes (Regions of Interest), and a feature matrix of size $\mathbb{R}^{n\times d}$, where d represents the feature dimension per node, for simplicity in analysis we assume that each GNN layer outputs embeddings of size d. Thus, the weight matrix for each layer is of size $\mathbb{R}^{d\times d}$. The orthonormalization step for each layer requires matrix multiplication, resulting in a complexity of $\mathcal{O}(Ld^3)$.

The forward and gradient steps of the algorithm have a complexity similar to that of a simple GNN, i.e., $\mathcal{O}(Lnd^2+Led)$, where e is the number of edges in the graph. Consequently, the total worst-case complexity of the proposed framework is $2\times\mathcal{O}(Lf^3 + Lnf^2 + Lef)$ (multiplication factor due to ensemble model).

It is worth noting that for fMRI analysis, the feature dimension d is typically of the same order as the number of nodes n. Therefore, the complexity can be simplified to $\mathcal{O}(Ln^3+Len)$. Hence, the worst-case complexity of our proposed algorithm remains in the same order as of GNN model.

Appendix C. Robustness Analysis: Noise in BOLD signal and demographic attributes

In analyzing the trustworthiness of the proposed framework, we introduced adversarial noise into the BOLD signals. Specifically, we added Gaussian noise with zero mean and variance σ^2. The results, presented in Figure C.1, demonstrate that the proposed

Table A.1: Comprehensive statistical overview of the REST-meta-MDD dataset utilized in this study.

Class	# of samples	Female	Male	Age	Education
HC	756	446	310	34.64 ± 13.17	13.56 ± 3.42
MDD	814	518	296	34.45 ± 11.61	11.95 ± 3.38
Total	1570	964	606	-	-

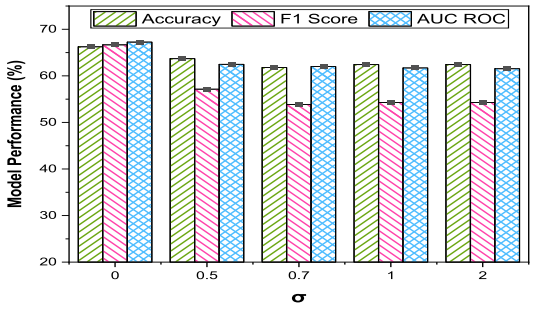


Figure C.1: Robustness analysis under addition of Gaussian noise in BOLD signal

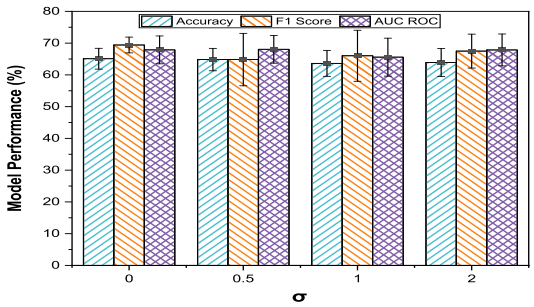


Figure C.2: Robustness analysis under addition of Gaussian noise in demographic attributes

framework consistently mitigates the effects of adversarial noise across varying standard deviations. This enhanced robustness is attributed to the intra- and inter-network diversity maximization employed in the ensemble network. Additionally, our framework shows adaptability to noise in demographic attributes, bolstered by a layer-wise weight orthonormalization process that increases resilience to feature perturbations. Robustness to noise in the Age attribute is further evaluated, with results displayed in Figure C.2.

Appendix D. Generalization of Framework on Different Task

We have demonstrated the flexibility of our proposed framework by evaluating it on an additional classification task within the same Rest-meta-MDD dataset: the classification of patients with Recurrent Major Depressive Disorder (rMDD) vs. HC. This task is highly relevant to the neuroscience community, as treatment and rehabilitation approaches differ for patients with recurrent depressive episodes than single episode depression (Cui et al., 2024; Otte et al.). The results of this evaluation are summarized in Table D.2, demonstrating the efficacy of the proposed framework in a different classification problem as well.

Appendix E. Sensitivity of Diversity Regularizer

To assess the sensitivity of the diversity regularizer β on the performance of CoRE-BOLD, we varied its value from 0.2 to 2. The corresponding results are presented in Figure E.3.

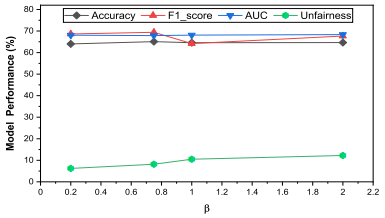


Figure E.3: Analysis of inter-network diversity.

Appendix F. Table of Notations

To improve the clarity of the proposed framework, we provide a concise summary of the key variables and notations used throughout the paper in Table F.3.

.

Table D.2: Classification of rMDD vs. Healthy Controls (HC)

Models	**Accuracy (↑)**	**AUC-ROC (↑)**	**SER (↓)**
GCN	69.11 ± 1.03	58.59 ± 1.21	1.35 ± 0.68
GIN	69.76 ± 0.95	59.05 ± 7.93	2.11 ± 2.39
GAT	68.78 ± 0.65	59.91 ± 6.08	1.18 ± 0.34
BrainGNN	64.55 ± 2.98	49.93 ± 1.02	1.16 ± 0.15
BrainNetCNN	65.85 ± 3.95	51.81 ± 5.07	3.17 ± 2.05
TGCN	66.27 ± 2.81	69.03 ± 1.56	1.10 ± 0.11
Weighted ensemble	72.36 ± 3.29	$\mathbf{73.29 \pm 2.60}$	1.14 ± 0.02
CoRE-BOLD	$\mathbf{76.12 \pm 2.94}$	72.52 ± 2.18	$\mathbf{1.04 \pm 0.03}$

Table F.3: Table to describe the notations used in the paper.

Notation	**Definition**	**Notation**	**Definition**
S_i	Subject i	$\mathbf{B}_i$	BOLD Signal of subject i
T_i	Time duration of BOLD Signal i	$\mathbf{A}_i$	Brain graph of subject i
$\mathbf{X}_i$	Feature matrix corresponding to A_i	$f_{\boldsymbol{\Theta}}(\cdot)$	GNN
y_i	Predicted label	$\boldsymbol{\Theta}$	GNN parameters
$\mathbf{C}_i$	ROI similarity matrix of subject i	$\mathbf{b}_i^{(k)}$	BOLD signal for k^{th} ROI in i^{th} subject
$\wp$	Pearson Correlation	Coh	Coherence measure
$\mathcal{D}$	Weighted diversity	$\delta(\cdot)$	Cosine similarity
$\mathcal{L}_{\text{div}}$	Inter-Network diversity loss	$\mathcal{L}_{\text{prejudice}}$	Unfairness loss
G_{0_i}, G_{1_i}	Groups for i^{th} sensitive attribute	$\mathbf{e}_u^{(k)}$	Embedding of node u after k layers
$\mathcal{N}_u$	Neighborhood of node u	$\mathbf{b}^{(k)}$	Bias parameters of GNN
$\mathcal{S}_{\boldsymbol{\Theta}}$	Set of matrices over Stiefel manifold	-	-