题目

摘要

开发基于潜在生物机制而非主观症状的新型精神类疾病诊断模型已成为共识。近年来，基于功能连接性（FC）的机器学习分类器被用于识别精神类疾病患者与健康个体的大脑生物标志物。然而，现有基于机器学习的诊断模型存在过拟合（例如，因训练样本不足导致）问题，且在新数据上表现较差。此外，难以获得可解释的、可靠的大脑生物标志物来揭示诊断决策背后的神经机制，这进一步限制了其在临床中的应用。本次实验探索了基于信息瓶颈的图神经网络（GNN）分类框架框架-BrainIB, 并验证了其在基于 fMRI 的精神疾病分类任务上的有效性。本文进一步尝试对该框架进行了优化，通过引入自注意力机制增强了模型在大脑网络中远距离脑区之间的依赖建模能力。实验结果表明，改进后的模型在各性能指标上均有一定提升。同时，本文探讨了未来通过结合多模态数据（如 EEG 和结构 MRI）进一步提高诊断性能的潜力。

关键词：精神类疾病; 图神经网络; 大脑网络; 功能磁共振成像

1 引言

精神障碍（如抑郁症和自闭症）是全球致残的主要原因，然而精神障碍的诊断仍然是一个具有挑战性的难题。传统的精神疾病诊断方法依赖于临床评估，主观性较强且由于临床症状个体差异较大，有可能会导致误诊。因此，基于潜在生物学机制而非仅依赖症状来开发有效的诊断工具，已成为当前医学研究领域的共识。这类方法的核心在于利用先进的神经影像学技术和生物标志物，为精神障碍的早期诊断和个性化治疗提供更加准确的依据。

功能性磁共振成像（fMRI）是一种非侵入性的神经影像技术，被广泛应用于揭示精神障碍的潜在病理生理机制。静息态 fMRI（rs-fMRI）能够研究和评估不同患者群体大脑功能连接网络（FC）的变化 [2, 27]。以往的研究表明，基于 fMRI 的表征可以作为现有症状诊断的可靠补充 [2, 27]。通过使用 fMRI 技术，研究人员能够捕捉到大脑不同区域之间的功能连接情况，为精神障碍的早期诊断和个性化治疗提供更为客观的生物标志物。

近年来，机器学习（ML）技术的迅速发展为基于 fMRI 数据的精神障碍诊断提供了新的视角和方法。传统的机器学习方法，如支持向量机（SVM） [16] 和随机森林（RF） [20]，已经广泛应用于 fMRI 功能连接性（FC）度量值的分类分析。然而，这些早期方法通常采用较为浅层或简单的分类模型，无法有效处理大脑网络中潜在的非线性关系和复杂性。

深度神经网络（DNN）近年来因其强大的表示能力和对复杂模式的捕捉能力而受到广泛关注 [11]。与传统的浅层学习模型不同，深度神经网络能够通过多个隐藏层学习数据的多层次特征，因此在大脑功能连接网络的复杂分析中表现出了显著优势。大脑网络通常可以被视

为复杂的图结构，其中大脑区域（ROIs）作为节点，功能连接性（FC）作为节点之间的边。这种图结构的特点激发了图神经网络（GNNs） [9] 在精神障碍诊断中的应用。图神经网络通过处理图结构数据，能够捕捉到大脑不同区域之间的复杂相互关系，极大地提高了分析和诊断的精度。

尽管近年来机器学习在精神障碍诊断中的应用取得了显著的性能提升，但现有的基于机器学习的诊断模型仍面临一些重要问题：

1. **可解释性**：大多数现有的诊断模型 [19, 28] 未能有效揭示可解释的大脑生物标志物（如将大脑区域（ROIs）视为一组节点或边），这些生物标志物有助于阐明诊断决策背后的机制并揭示疾病的神经机制。缺乏可解释性使得这些模型的临床应用受到限制，尤其是在需要临床医生理解和验证模型决策的情境中。

2. **泛化能力**：大多数现有的诊断模型在同质化或单一数据集上进行训练，并且样本量通常较小。这种训练方式容易导致过拟合，从而使得模型在实际部署过程中难以有效地处理不同来源的数据。例如，某些研究 [32] 仅考虑了 24 名患有重度抑郁症的患者，这样小规模的样本和单一数据来源限制了模型的泛化能力。

信息瓶颈（IB）原则 [24] 为模型可解释性和泛化能力问题的研究提供了方向。信息瓶颈原则源自信息理论，旨在从原始数据中提取出最具信息量的紧凑表示，以便更有效地进行标签预测。通过消除冗余信息，这种方法能够显著增强模型的泛化能力并提高其可解释性 [22]。近年来，信息瓶颈的思想已被扩展到图神经网络（GNNs）中，以更好地处理图结构数据。然而，当前的图信息瓶颈方法（如子图信息瓶颈（SIB） [29]）通常是基于节点选择而非边选择来识别重要子图，这在精神障碍诊断中存在一定的局限性。在精神障碍的诊断中，大脑的功能连接（边）比单纯的大脑区域（节点）更加重要。首先，越来越多的证据表明，大脑的连接性在精神障碍的诊断和神经机制理解中起着至关重要的作用 [8, 12]。例如，Insel 等人 [8] 在《科学》杂志中提出，未来的诊断可能将“精神障碍”重新定义为“脑回路障碍”。这一观点强调了大脑各区域之间的相互连接，而非单一的脑区本身，是理解和诊断精神障碍的关键。其次，许多针对精神障碍的诊断研究已发现了与大脑连接相关的潜在生物标志物 [5, 10]，这些研究进一步证明了边（即连接）在精神健康领域诊断中的重要性。

尽管图信息瓶颈方法在小规模数据集上取得了一定的成果 [4]，这些方法的性能和适用性在大规模数据集中的应用仍未得到充分验证。特别是在涉及到大脑疾病的大规模数据集时，这些数据集通常包含数百个节点和成千上万条边，现有的图信息瓶颈方法尚未能充分展示其在如此复杂和大规模数据上的有效性。因此，如何将信息瓶颈方法有效地扩展到大规模脑连接网络数据，仍然是未来研究的重要挑战。

本文探索了 zheng 等人 [1] 开发的 BrainIB 模型，模型引入了信息瓶颈（IB）原则来分析大脑网络，并基于图神经网络（GNN）构建了一个可解释的大脑网络分类框架。该框架能够识别与决策最相关的子图，并且在未见过的数据上具有良好的泛化能力。

2 相关工作

2.1 精神疾病诊断模型

在现代精神障碍诊断模型的开发中，识别预测性子网络和连接（边）是一个至关重要的步骤 [25]。传统的方法通常将功能连接性作为特征，采用特征选择技术来保留最显著的连接。常见的特征选择方法包括统计检验 [23]，如 t 检验或秩和检验，以及 LASSO（最小绝对收缩和选择算子）。这些方法通过选择对分类最有用的特征来提高模型的性能。然而，这些方法通常依赖于人工设计和显式的特征选择过程，可能无法充分利用数据中潜在的复杂关系。

早期的精神障碍分类模型主要采用传统的机器学习方法，如支持向量机（SVM）和随机森林（RF）人 [16]。例如，Plitt 等人 [18] 使用线性 SVM 对自闭症患者和健康对照组进行区分，达到了 0.697 的整体准确率。然而，这些浅层学习方法无法捕捉复杂大脑网络结构中的拓扑信息 [7]，因此在大规模数据集上的表现往往不足以满足临床需求。为了进一步提高诊断准确性，最新的研究开始转向图神经网络（GNNs）。例如，Parisot 等人 [17] 在自闭症数据集上应用了图卷积网络（GCNs），并取得了更高的准确率（0.704）。尽管 GNNs 在性能上取得了明显的提升，但它们通常被认为是“黑箱”算法，缺乏透明的决策过程，这对于临床应用来说是一个重大问题。

2.2 信息瓶颈以及 GNN 可解释性

信息瓶颈（IB）原则最近被扩展到图神经网络（GNNs）中。给定一个图数据 G，它同时编码了图的结构信息（通过邻接矩阵 A 来表示）和节点特征矩阵 X，以及标签 Y。子图信息瓶颈（SIB） [29] 旨在通过以下目标函数从图 G 中提取最具信息量或可解释性的子图 Gsub：

$$\mathcal{L}_{\text{SIB}} = \min I(G; G_{\text{sub}}) - \beta I(Y; G_{\text{sub}})$$

Yu 等人 [29] 通过最小化交叉熵损失来近似 –I(Y;Gsub)，从而提取出对诊断决策至关重要的子图。该方法通过去除冗余或无关的节点来实现子图的选择。其中互信息项 I(G;Gsub) 通过互信息神经估计器（MINE） [1] 进行评估，但 MINE 需要额外的网络，并且在训练过程中可能会变得不稳定。

SIB 可以被视为一种内置的可解释图神经网络（即自解释 GNNs），它能够自动识别对决策或图标签 Y 影响最大的子图。图神经网络的可解释性在近期得到了广泛的关注，然而，大多数现有的解释方法是后置（post-hoc）的，即这些方法通过另一个解释性模型来为已训练好的 GNN 提供解释 [13, 31]。与自我解释方法相比，事后解释在潜在诊断决策过程中是否可靠仍然是一个问题 [21]。在大脑网络分类应用中，BrainNNExplainer [3] 是一个基于图神经网络的可解释模型，它通过学习全局掩码来突出疾病特异性的显著大脑网络连接。另一个例子是 BrainGNN [11]，它设计了 ROI 感知的图卷积层和池化层，用于突出与疾病相关的显著大脑区域（即图中的节点）。

3 本文方法

3.1 问题定义

给定一组加权大脑网络 $\{G^1, G^2, \ldots, G^N\}$，模型的输出对应标签为 $\{y^1, y^2, \ldots, y^N\}$。对于第 i 个大脑网络 $G^i = (A^i, X^i)$，其中 A^i 是用于描述图结构的邻接矩阵，$A^i \in \{0,1\}^{n\times n}$，而 X^i 是由加权功能连接性数值构成的节点特征矩阵，$X^i \in \mathbb{R}^{n\times n}$。具体来说，$A^i$ 是一个二值化的功能连接（FC）矩阵，其中仅保留相关性绝对值排名前 20% 的连接；将其转换为 1，而其他连接则转换为 0。对于节点特征矩阵 X，节点 k 的特征向量 x_k^i 定义为：$x_k^i = [p_{k1}, \ldots, p_{kn}]^T$。其中，$p_{kl}$ 是节点 k 和节点 l 之间的皮尔逊相关系数。

3.2 本文方法概述

BrainIB 的流程图如图 1所示。BrainIB 包括三个模块：子图生成器、图编码器和互信息估计模块。子图生成器用于从原始图 G 中采样子图 G_{sub}。图编码器用于从 G 或 G_{sub} 中学习图嵌入。互信息估计模块用于评估 G 或 G_{sub} 的互信息。

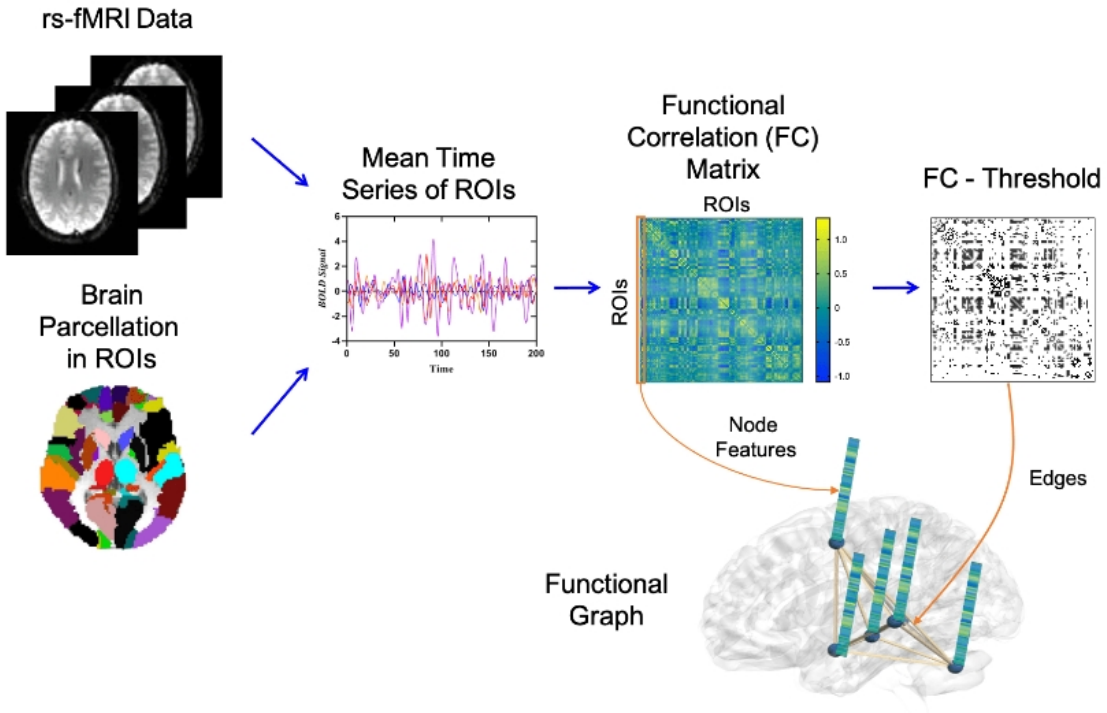


图 1. 方法示意图

3.3 子图生成器

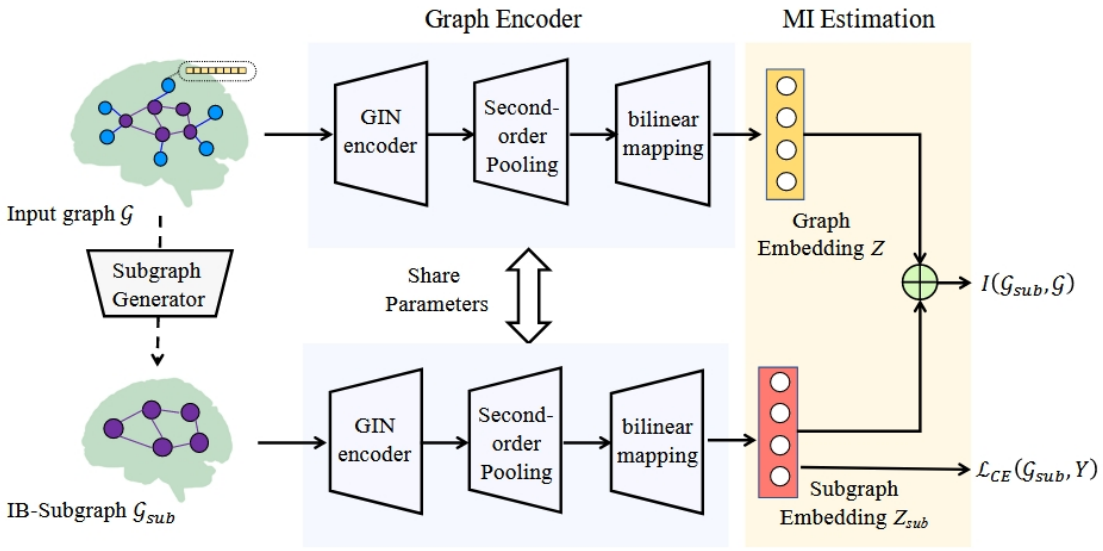


图 2. 子图生成器

子图生成模块的过程如图 2所示。BranIB 通过边分配而非节点分配从输入图中生成 IB-子图（IB-Subgraph）。其中，边分配和节点分配分别表示图中的每条边和节点是否被包含在 IB 子图中。对于给定的图 G=(A,X)，模型通过节点特征 X 计算边是否被包含在子图中的概率，从而确定边分配矩阵 S。

首先，模型会学习一个边注意力掩码 P，其中每个元素表示边的选择概率。为简化计算，节点特征 X 直接被输入到一个多层感知机（MLP），随后通过 Sigmoid 函数获得边注意力掩码 $P \in \mathbb{R}^{n\times n}$ 并确保 $p_{ij} \in [0,1]$，其中 n 是节点数量，p_{ij} 表示节点 i 和节点 j 之间的边被选择的概率。具体而言，p_{ij} 可以被定义为：

$$p_{ij} = \sigma\left(X_i W_j^T\right),$$

其中，$\sigma(\cdot)$ 是 Sigmoid 函数，$X_i = [p_{i1}, \ldots, p_{in}]$，$p_{ij}$ 是节点 i 和节点 j 的皮尔逊相关系数，W_j^T 表示 MLP 每一层中第 j 个神经元的参数。这表示节点 i 的特征通过 MLP 的第 j 个神经元输出，计算出连接节点 i 和节点 j 的边的概率 p_{ij}。

其中，$\sigma(\cdot)$ 是 Sigmoid 函数，$X_i = [p_{i1}, \ldots, p_{in}]$，$p_{ij}$ 是节点 i 和节点 j 的皮尔逊相关系数，W_j^T 表示 MLP 每一层中第 j 个神经元的参数。这表示节点 i 的特征通过 MLP 的第 j 个神经元输出，计算出连接节点 i 和节点 j 的边的概率 p_{ij}。

接着，将 P 二值化以获得边分配 $S \in \{0,1\}^{n\times n}$。为了确保梯度关于 p_{ij} 可计算，采用 Gumbel-Softmax 重参数化技巧 [14] 来更新边分配 S。通常情况下，Gumbel-Softmax 将一个连续随机变量近似为独热向量，但直接应用于 P 时，它只能为连续 n 条边选择一条。为保留足够数量的边，P 被重塑为 K 维矩阵，然后通过 Gumbel-Softmax 方法进行二值化，从而生成边分配 S。这种方法确保在 K 条连接边中至少保留一条边。具体来说，K 维样本向量中第 k 条边的采样概率定义为：

$$\hat{p}_k = \frac{\exp\left((\log p_k + c_k)/\tau\right)}{\sum_{i=1}^{K} \exp\left((\log p_i + c_i)/\tau\right)},$$

其中 τ 是 Concrete 分布的温度参数，p 是边的选择概率，$\hat{p}$ 是样本概率。c_k 从 Gumbel$(0,1)$ 分布中生成：

$$c_k = -\log\left(-\log U_k\right), \quad U_k \sim \text{Uniform}(0,1)$$

通过这一过程，K 决定了 IB 子图 G_{sub} 的规模，最终能够保留 $n \times n/K$ 条边。接下来，S 被转换为 $n \times n$ 矩阵，IB 子图 G_{sub} 可通过 $A_{\text{sub}} = A \odot S$ 提取，其中 $\odot$ 表示逐元素乘法；A 和 A_{sub} 分别表示输入图 G 的邻接矩阵和 IB 子图 G_{sub} 的邻接矩阵。

3.4 图编码器

图编码器模块由 GIN 编码器和双线性映射二阶池化 [26] 组成。经过 GIN 编码器后，从原始节点特征矩阵 X 中获得节点表示 $H \in \mathbb{R}^{n \times d}$。接着，应用双线性映射二阶池化方法，从 H 中生成向量化的图嵌入。与现有的图池化方法相比，双线性映射二阶池化能够利用所有节点的信息，收集二阶统计量，并有效减少训练参数的数量 [26]。

给定节点表示 $H = [h_1, h_2, \ldots, h_n] \in \mathbb{R}^{n \times d}$ 和 $W \in \mathbb{R}^{d \times d'}$（表示从 $\mathbb{R}^d$ 到 $\mathbb{R}^{d'}$ 的可训练线性映射矩阵），双线性映射二阶池化（$\text{SOPOOL}_{\text{bimap}}$）的定义为：

$$\text{SOPOOL}_{\text{bimap}}(H) = \text{SOPOOL}(HW) = W^T H^T H W \in \mathbb{R}^{d' \times d'}$$

直接将 $\text{SOPOOL}_{\text{bimap}}$ 应用于 $H = \text{GIN}(A, X)$，然后将输出矩阵展平为一个 d'^2 维的图嵌入向量 h_g：

$$h_g = \text{FLATTEN}\left(\text{SOPOOL}_{\text{bimap}}(\text{GIN}(A, X))\right) \in \mathbb{R}^{d'^2}$$

3.5 互信息估计

图 IB 目标函数包含两个互信息项，具体形式如下：

$$\min_{G_{\text{sub}}} I(G_{\text{sub}}, G) - \beta I(G_{\text{sub}}, Y)$$

其中，最小化 $-I(G_{\text{sub}}, Y)$（即最大化 $I(G_{\text{sub}}, Y)$）是为了鼓励子图 G_{sub} 对图标签 Y 具有更强的可预测性。从数学上来看，有：

$$-I(G_{\text{sub}}, Y) \leq \mathbb{E}_{Y_{\text{sub}}} - \log q_\theta(Y|G_{\text{sub}}) := \mathcal{L}_{CE}(G_{\text{sub}}, Y),$$

其中 $q_\theta(Y|G_{\text{sub}})$ 是从 G_{sub} 到 Y 的映射的变分近似，公式表明最小化 $-I(G_{\text{sub}}, Y)$ 约等于最小化交叉熵损失 $\mathcal{L}_{CE}$。

对于互信息项 $I(G_{\text{sub}}, G)$，首先通过图编码器分别从输入图 G 和子图 G_{sub} 中提取嵌入 Z 和 Z_{sub}。根据有效编码假设 [23]，Z 在编码过程中是无信息损失的，因此将 $I(G_{\text{sub}}, G)$ 近似为 $I(Z_{\text{sub}}, Z)$。与使用 MINE 需要额外神经网络的 SIB 方法不同，BrainIB 采用最近提出的基于矩阵的 Rényi' s α 阶互信息估计 [6,30]。这种方法在数学上定义明确且计算高效。

具体来说，给定一个大小为 N 的小批量样本，可分别获得 $\{Z_i\}_{i=1}^N$ 和 $\{Z_{\text{sub},i}\}_{i=1}^N$，其中 Z_i 和 Z_{sub} 分别表示第 i 个图和子图的图嵌入。根据 [6]，可以通过样本的（归一化后的）Gram 矩阵 D 的特征谱来估计图嵌入的熵：

$$H_\alpha(Z) = \frac{1}{1-\alpha} \log_2\left(\text{tr}\left(D^\alpha\right)\right) = \frac{1}{1-\alpha} \log_2\left(\sum_{i=1} \lambda_i(D)^\alpha\right),$$

其中 tr 表示矩阵的迹，$D = K/\mathrm{tr}(K)$，$K = \kappa(Z_i, Z_j)$ 是通过正定核函数计算得到的 Gram 矩阵，λ_i 是 D 的第 i 个特征值。在极限情况下，当 $\alpha \to 1$ 时，上式会收敛到与 Shannon 熵类似的度量 $H(Z)$。

类似地，可以通过 $\{Z_{\mathrm{sub},i}\}_{i=1}^{N}$ 估计子图嵌入的熵：

$$H_\alpha(Z_{\mathrm{sub}}) = \frac{1}{1-\alpha}\log_2\left(\mathrm{tr}\left(D_{\mathrm{sub}}^{\alpha}\right)\right) = \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}\lambda_i(D_{\mathrm{sub}})^{\alpha}\right),$$

其中 D_{sub} 是使用子图嵌入计算得到的 Gram 矩阵。

图嵌入 Z 和子图嵌入 Z_{sub} 的联合熵可以通过以下公式估计：

$$H_\alpha(Z, Z_{\mathrm{sub}}) = H_\alpha\left(\frac{D \circ D_{\mathrm{sub}}}{\mathrm{tr}(D \circ D_{\mathrm{sub}})}\right),$$

其中 $D \circ D_{\mathrm{sub}}$ 表示 D 和 D_{sub} 的 Hadamard 乘积。

根据前面几个公式，基于矩阵的 Rényi' s α 阶互信息 $I(Z_{\mathrm{sub}}, Z)$ 类似于 Shannon 互信息的定义 [30]，具体为：

$$I(Z_{\mathrm{sub}}, Z) = H_\alpha(Z_{\mathrm{sub}}) + H_\alpha(Z) - H_\alpha(Z_{\mathrm{sub}}, Z)$$

BrainIB 使用径向基函数（RBF）核来计算 D 和 D_{sub}。具体为：

$$\kappa(z^i, z^j) = \exp\left(-\frac{\|z^i - z^j\|^2}{2\sigma^2}\right),$$

其中核宽度 σ 通过每个样本的 $k(k = 10)$ 个最近邻的距离均值估计，并对所有样本取平均值来确定。此外，α 被固定为 $\alpha = 1.01$ 来近似 Shannon 互信息。

最终，BrainIB 的目标函数可以表示为：

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(G_{\mathrm{sub}}, Y) + \beta I(G_{\mathrm{sub}}, G)$$

其中, β 是超参数

4 复现细节

4.1 数据集介绍

本文使用了 The Autism Brain Imaging Data Exchange(ABIDE) 数据集。该数据集的人口统计学与临床特征信息如表1所示

特征	ASD	TD
样本大小	528	571
年龄（岁）	17.0 ± 8.4	17.1 ± 7.7
性别（男/女）	464/64	471/100

表 1. 人口统计学和临床特征信息

ABIDE 数据集 [15] 是汇集并公开共享了超过 1,000 份现有的静息态功能磁共振成像（rs-fMRI）数据。其中 ABIDE-I 提供了 528 名自闭症谱系障碍（ASD）患者和 571 名正常发育（TD）个体的数据。本次实验通过统计参数映射（SPM）软件对静息态 fMRI 原始数据进行了预处理。具体预处理步骤如下：首先，对静息态 fMRI 图像进行切片时间校正，以补偿不同切片采集时间的差异。然后，进行重对齐（realignment），通过平移和旋转校正不同时间点 fMRI 图像间由于头部运动产生的位移误差。接着，使用从个体 fMRI 图像到 MNI（Montreal Neurological Institue）模板的变形参数，将静息态 fMRI 图像标准化到标准空间。随后，使用半高宽为 6 mm 的高斯滤波器对功能图像进行空间平滑。最后，对处理后的 fMRI 图像使用带通滤波器（0.01–0.08 Hz）进行时间滤波，并回归掉头部运动、白质信号和脑脊髓液（CSF）信号的影响。

4.2 创新点

在 BrainIB 框架中，图编码器主要通过 GIN 来提取局部邻域信息，但在捕获大脑网络中的长程依赖关系方面存在局限。由于精神类疾病的诊断依赖于脑区间的全局交互，单靠局部特征可能无法充分表征大脑网络结构。因此在探索模型的同时，本次实验尝试引入了 Transformer 自注意力机制，直接建模大脑网络中远距离脑区间的依赖关系，增强模型对全局信息的整合能力。与此同时，Transformer 能够自动关注最相关的脑区连接，进一步提升模型的泛化能力和分类表现。自注意力机制被引入到图编码器模块的节点表示生成阶段。在 GIN 提取节点局部邻域特征后，节点表示通过 Transformer 进一步建模全局依赖关系，捕获大脑网络中远距离脑区之间的交互信息。Transformer 的输出随后经过池化生成图级嵌入，用于最终分类任务。

5 实验结果分析

5.1 十折交叉验证

本次实验通过十折交叉验证评估模型在 ABIDE 数据集上的性能，并以准确率、F1 分数以及 Matthews 相关系数（MCC）作为评估指标。表 2 展示了 brainIB 模型在 10 折交叉验证中的性能。

	ACC	F1	MCC
原文	0.700	-	-
复现	0.681	0.744	0.454

表 2. BrainIB 模型在 ABIDE 数据集上的十折交叉验证性能

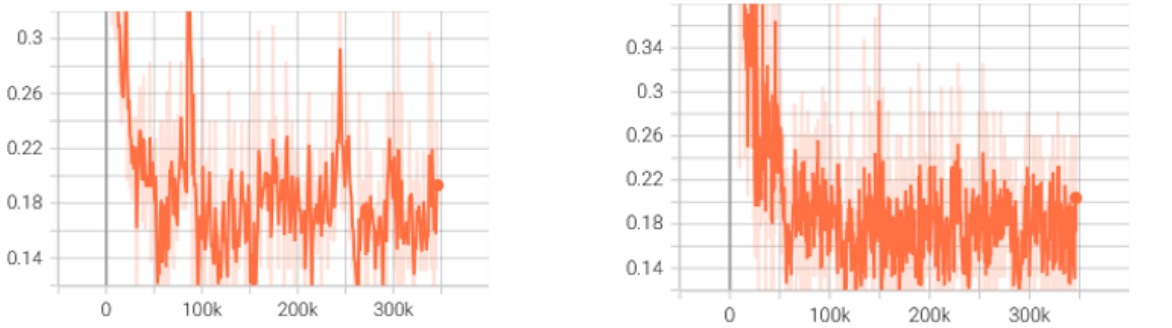
本次实验复现的结果准确率为 0.681，略低于原文，但分类性能依旧表现良好。同时，由于在精神类疾病诊断任务中，数据通常存在不平衡问题，正类样本（如患者）较少，导致准确率（ACC）可能无法反映模型的真实性能。为此，本次实验增加了 F1 分数和 MCC 作为补充指标：F1 分数衡量了查准率和查全率，适合评估正类识别能力，而 MCC 能全面衡量模型在不平衡数据下的分类表现。本次复现结果中 F1 分数达到了 0.744，表明模型在查准率和查全率方面取得了良好的平衡，表现出较好的正类识别能力。MCC 为 0.454，说明模型在不平衡数据下对正负类的分类具有一定的区分能力，但仍有优化空间。

表3展示了引入 transformer 后的性能指标。可以看出，引入自注意力机制后，模型的性能要略高于复现结果但要准确率要略低于论文中的结果。

	ABIDE		
	ACC	F1	MCC
With Transformer	0.694	0.752	0.462

表 3. 引入自注意力机制后，模型在 ABIDE 数据集上的十折交叉验证性能

图3给出了引入 transformer 前后模型损失变化的对比。可以看到，引入 Transformer 后，模型在训练初期能够更快地收敛。此外，在引入 transformer 后，模型训练损失的波动出现了减少，有效提升了模型训练的稳定性。



(a) 引入transformer前的训练损失变化　　(b) 引入transformer后的训练损失变化

图 3. 引入 transformer 前后模型训练损失变化对比

6 总结与展望

本文探索了一种基于图信息瓶颈方法的脑网络诊断模型，通过结合图神经网络和信息理论，有效提升了精神疾病诊断的泛化能力和可解释性。复现结果表明，该模型在自闭症数据集上的性能较好，并通过引入自注意力机制进一步优化了对远距离脑区交互信息的捕获。

fMRI 具有较高的空间分辨率，能够较为精确描绘大脑区域的活动分布，但其时间分辨率较低（通常为秒级），难以捕捉快速变化的脑活动。因此，未来可以结合 EEG 等高时间分辨率的数据，更精准地刻画大脑的动态功能连接特征。此外，结构 MRI 和 DTI 等模态也可补充大脑解剖和白质纤维的结构信息，从多个层面全面表征大脑特征。利用多模态数据来弥补 fMRI 单一模态的不足，进一步提升模型的泛化能力和诊断准确性。

参考文献

[1] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

[2] Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Stephen M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

[3] H. Cui, W. Dai, Y. Zhu, X. Li, L. He, and C. Yang. Brainnnexplainer: An interpretable graph neural network framework for brain network based disease analysis. In *International Conference on Machine Learning*, 2021.

[4] A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797, 1991.

[5] S. Gallo, A. El-Gazzar, P. Zhutovsky, R. M. Thomas, N. Javaheripour, M. Li, L. Bartova, D. Bathula, U. Dannlowski, C. Davey, et al. Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies. *Molecular Psychiatry*, pages 1–10, 2023.

[6] L. G. S. Giraldo, M. Rao, and J. C. Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

[7] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan. Identifying autism spectrum disorder from resting-state fmri using deep belief network. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):2847–2861, 2020.

[8] T. R. Insel and B. N. Cuthbert. Brain disorders? precisely. *Science*, 348(6234):499–500, 2015.

[9] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[10] A. Li, A. Zalesky, W. Yue, O. Howes, H. Yan, Y. Liu, L. Fan, K. J. Whitaker, K. Xu, G. Rao, et al. A neuroimaging biomarker for striatal dysfunction in schizophrenia. *Nature Medicine*, 26(4):558–565, 2020.

[11] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

[12] A. M. Lozano. Waving hello to noninvasive deep-brain stimulation. *New England Journal of Medicine*, 377(11):1096–1098, 2017.

[13] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, volume 33, pages 19620–19631, 2020.

[14] C. Maddison, A. Mnih, and Y. Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the International Conference on Learning Representations*, 2017.

[15] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014.

[16] X. Pan and Y. Xu. A novel and safe two-stage screening method for support vector machine. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8):2263–2274, 2018.

[17] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer' s disease. *Medical Image Analysis*, 48:117–130, 2018.

[18] M. Plitt, K. A. Barnes, and A. Martin. Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *Neuroimage Clinical*, 7(C), 2015.

[19] Z. Rakhimberdina, X. Liu, and T. Murata. Population graph-based multi-model ensemble method for diagnosing autism spectrum disorder. *Sensors*, 20(21):6001, 2020.

[20] J. F. A. Ronicko, J. Thomas, P. Thangavel, V. Koneru, G. Langs, and J. Dauwels. Diagnostic classification of autism using resting-state fmri data improves with full correlation functional brain connectivity compared to partial correlation. *Journal of Neuroscience Methods*, 345:108884, 2020.

[21] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[22] O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

[23] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

[24] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communications, Control, and Computing*, pages 368–377, 1999.

[25] L. Wang, F. V. Lin, M. Cole, and Z. Zhang. Learning clique subgraphs in structural brain network classification with application to crystallized cognition. *Neuroimage*, 225:117493, 2021.

[26] Z. Wang and S. Ji. Second-order pooling for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[27] M. Xia and Y. He. Functional connectomics from a "big data" perspective. *Neuroimage*, 160:152–167, 2017.

[28] D. Yao, M. Liu, M. Wang, C. Lian, J. Wei, L. Sun, J. Sui, and D. Shen. Triplet graph convolutional network for multi-scale analysis of functional connectivity using functional mri. In *International Workshop on Graph Learning in Medical Imaging*, pages 70–78. Springer, 2019.

[29] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He. Recognizing predictive substructures with subgraph information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[30] S. Yu, L. G. S. Giraldo, R. Jenssen, and J. C. Principe. Multivariate extension of matrix-based renyi' s α-order entropy functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2960–2966, 2019.

[31] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji. On explainability of graph neural networks via subgraph explorations. In *International Conference on Machine Learning*, pages 12241–12252. PMLR, 2021.

[32] L.-L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li, and D. Hu. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain*, 135(5):1498–1507, 2012.