arXiv:2603.02518v1 [cs.CV] 3 Mar 2026

超越解剖学：通过功能分区和图注意力网络对静息态功能磁共振成像进行可解释的ASD分类

SyedaHareem Madani[1], Noureen Bibi[1], Adam Rafiq Jeraj[2], Sumra Khan[1],Anas Zafar[3],和 Rizwan Qureshi[1]

[1] Salim Habib University[2] VentureDive[3] 计算机与新兴科学国立大学

摘要。 基于静息态功能磁共振成像的自闭症谱系障碍分类主要采用脑解剖分区，但其边界僵化，可能无法捕捉自闭症谱系障碍特有的分布式连接模式。我们提出一种基于图的深度学习框架，通过在 ABIDE I 数据集上比较解剖分区策略（AAL，116 个感兴趣区域）与功能分区策略（MSDL，39 个感兴趣区域），系统地解决这一局限。我们采用严格的手动 FSL 预处理流程处理 400 名受试者（200 名自闭症谱系障碍受试者、200 名典型发育对照者）的多中心异质性数据，并按站点分层，以 70/15/15 的比例划分数据，防止数据站点泄漏。高斯噪声增强仅应用于训练折，在保留留出评估集完整性的同时，将训练样本从 280 个扩增至 1,680 个。三阶段流程从采用 AAL 的基线图卷积网络（准确率 73.3%，AUC=0.74）开始，继而采用 MSDL 优化图卷积网络（84.0%，AUC=0.84），最终优化得到图注意力网络集成，测试准确率达到 95.0%（AUC=0.98），优于 ABIDE I 上所有近期的基于图神经网络的基准。仅通过脑图谱替换获得的 10.7 个百分点提升表明，功能分区是该流程中影响最大的单一建模决策。基于梯度的显著性分析与 GNNExplainer 分析结果一致，均指向后扣带皮层和楔前叶这两个默认模式网络核心枢纽，验证模型决策反映的是已知的自闭症谱系障碍神经病理学，而非站点或采集伪影。我们的框架提供了一套可复现、可解释且方法严谨的流程，连接了图神经网络架构与临床神经科学。为支持研究复现和社区采用，所有代码、预处理脚本和处理后的图数据集将在论文录用后公开发布。

关键词： 自闭症谱系障碍 · 图注意力网络 · 功能分区 · 静息态功能磁共振成像 · 可解释人工智能 · ABIDE

1 引言

自闭症谱系障碍（ASD）是一种神经发育障碍，其特征是社交沟通障碍以及受限、重复的行

为 [1]。临床诊断依赖行为评估和发育史，这些主观过程往往导致诊断延迟，症状出现三年后才确诊 [2], ，从而限制了在关键发育窗口期进行早期干预的机会 [3]。静息态功能磁共振成像（静息态功能磁共振成像）测量内在的BOLD信号波动，揭示了与自闭症谱系障碍相关的异常连接，包括局部过度连接和长程连接不足 [4] ；默认模式网络和显著性网络的异常与社会认知障碍相关 [5, 6]。重要的是，自闭症谱系障碍个体间的内在连接存在显著差异，这种现象被称为特异性大脑 [7], ，给通用生物标志物发现带来了根本性挑战。

利用ABIDE I数据库 [8] 的深度学习方法，包括自编码器 [9], 、卷积网络 [10],和图神经网络 [11]，在自闭症谱系障碍分类方面已展现出前景。然而，这些方法存在一个关键局限：几乎普遍依赖自动解剖标记（AAL）图谱 [12],，其静态的、由结构定义的边界，对于一种以功能网络失调为特征的病症而言，造成了根本性的表征不匹配。图注意力网络 [13]能够以自适应重要性权重对连接性进行建模，但即便最具表达力的架构也会受其输入表征的制约。功能衍生图谱，如MSDL [14]，基于功能一致性定义脑区，更适合自闭症谱系障碍分布式、个体特异的失调 [7, 15] ，但在统一的图神经网络框架内尚无对照比较。

我们提出一个三阶段图神经网络框架，用以分离图谱选择的影响，并提供首个对照证据，表明功能分区是基于静息态功能磁共振成像的自闭症谱系障碍分类中影响最大的单一建模决策。我们的贡献如下：

– **受控脑图谱比较：** 在统一的图神经网络框架内首次系统比较AAL与MSDL，表明仅通过脑图谱替换即可获得10.7个百分点的准确率提升。– **严格的评估方案：** 在全部17个ABIDE站点上进行按站点分层的70/15/15划分，且数据增强严格限定于训练折。– **最先进的分类：** GAT 集成在ABIDE I上达到95.0%的准确率和AUC = 0.98，优于近期所有基于图神经网络的基准。– **经生物学验证的可解释性：** 汇聚的GNNExplainer与显著性分析识别出后扣带皮层和楔前叶这两个核心默认模式网络枢纽，与既有的自闭症谱系障碍神经病理学一致。

2 方法学

2.1 数据采集与参与者

从 ABIDE I数据库 中选取了 400 名受试者 [8], ，其中包括 200 名自闭症谱系障碍受试者和 200 名典型发育（TD）对照组受试者，并在 17 个采集站点间进行了均衡分配（年龄范围为 7–64 岁，自闭症谱系障碍组以男性为主）。

数据按站点分层，依照 70/15/15 的训练集/验证集/测试集划分进行分组，分别得到 280 名训练受试者、60 名验证受试者和 60 名测试受试者。人口统计学信息见表1。

表1：数据集的人口统计分布

特征	ASD	TD	总计
人数	200	200	400
性别分布	以男性为主	均衡	-
年龄范围（岁）	7–64	7–64	-
数据模态	静息态功能磁共振成像（4D）	静息态功能磁共振成像（4D）	-

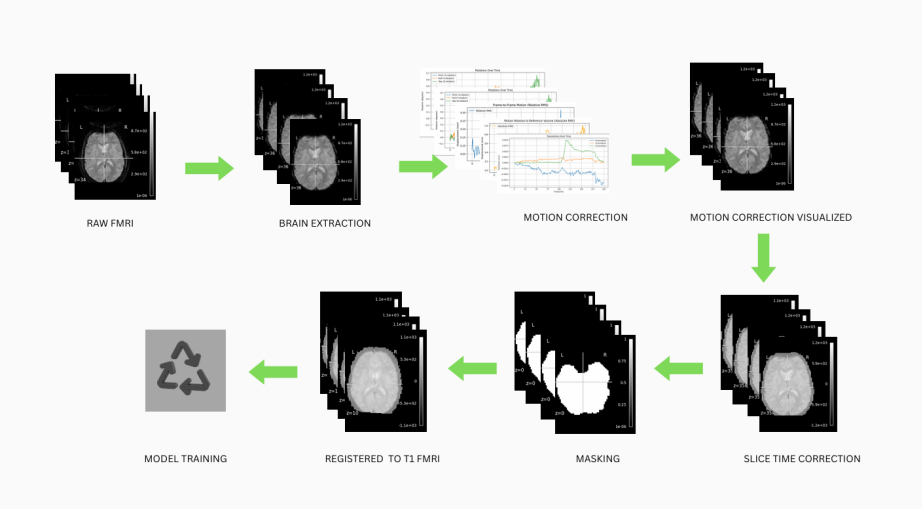
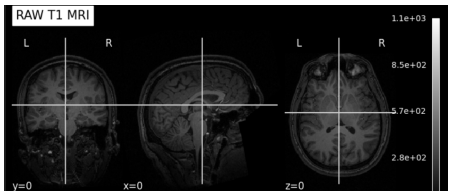


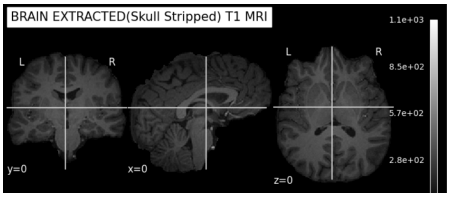
图1：完整的功能磁共振成像预处理流程。该流程从原始数据采集开始，依次经过标准的 FSL 预处理步骤、特征提取，最终进入模型训练。

2.2 结构预处理

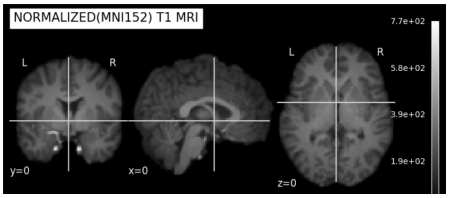
T1加权图像使用 FSL 的 BET（分数强度阈值为 0.5）进行去颅骨处理，随后通过仿射配准（FLIRT，12 个自由度）和非线性配准（FNIRT）至 MNI152 2毫米标准空间，如图 1 所示。目视检查确认所有受试者均成功完成配准。



(a) 去颅骨解剖图像，使用 BET



(b) 运动校正参数估计-由 MCFLIRT 估计



(c) 平均功能的配准体积到 T1

图 2：FSL结构预处理流程的可视化结果

2.3 功能预处理

功能预处理包括切片时序校正（FSL slicetimer，交错采集）、通过 MCFLIRT 进行运动校正（平均位移超过 3 毫米的体积被排除），以及使用 FLIRT（线性，12 个自由度）和 FNIRT（非线性）将图像空间标准化至 MNI152 2毫米标准空间。fMRI 到 T1 的配准采用基于边界的配准（BBR）。目视检查确认预处理成功，皮层边界得以保留且伪影极少（图 2）。

2.4 图构建与Atlas选择 ion

我们探索了两种不同的脑区划分策略，以评估解剖学可解释性与功能一致性之间的权衡。为进行生物学验证，我们使用自动解剖标记（AAL）图谱（116 个感兴趣区域）提取时间序列数据。通过对所有成对感兴趣区域的时间序列进行皮尔逊相关分析，生成了功能连接矩阵。为确保生成的图既符合生物学合理性，又便于计算处理，我们采用比例阈值化，仅保留最强的 20% 连接（图 3）。这一过程生成了一个包含 116 个节点的稀疏加权图表示，其中每个节点对应一个定义明确的解剖区域，有助于后续对模型决策进行神经生物学解释。

为了解决刚性解剖边界的局限性，并更好地捕捉与自闭症谱系障碍相关的异质性和个体特异性功能连接模式，优化后的模型采用了多被试字典学习

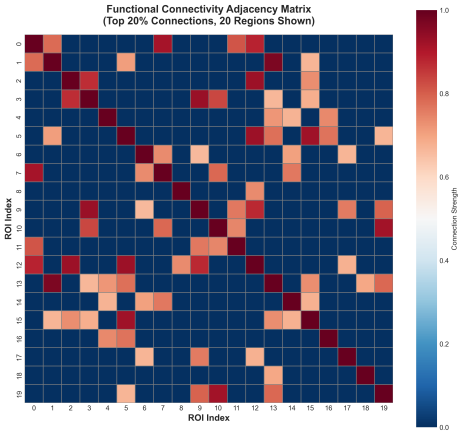


图 3：功能连接邻接矩阵。颜色越暖，表示脑区之间的时间相关性越强

（MSDL）Atlas。该Atlas基于受试者内部的功能一致性定义了39个概率感兴趣区，而非固定的解剖标志。脑网络同样被建模为一个图，其中节点表示这些以概率方式定义的功能边，网络则根据其各自时间序列中的皮尔逊相关强度进行加权。这种以功能为基础的表征在理论上更适合识别自闭症谱系障碍所特有的分布式且相互重叠的网络失调特征。

2.5 数据增强策略——高斯噪声注入

为应对标注医学数据稀缺的问题 [16]，我们实施了高斯噪声注入策略。在此过程中，我们通过添加随机高斯噪声（$\mu = 0$，$\sigma = 0.05$），为每位受试者的相关矩阵生成了 5 个增强版本。这使有效数据集规模从 400 个扩展至 2000 个图，从而显著增强了模型抵御过拟合的能力。

2.6 深度学习架构

评估了两种图神经网络架构，汇总见表2。基线GCN由两个图卷积层（$116 \to 64 \to 32$ 个单元）组成，采用 ReLU 激活函数和全局平均池化，聚合邻域信息以进行图级分类。GAT 集成采用两个注意力层，每层有两个注意力头（64 个隐藏通道，随机失活率为 0.5），并使用全局平均池化；在增强数据上训练五个独立模型，并通过软投票进行组合。训练期间还应用 DropEdge（丢弃 20% 的边），以防止过拟合。所有模型均使用 Adam（$\mathrm{lr} = 1\text{e-}3$）在 NVIDIA RTX 1630 GPU 上训练 100 个周期（总计 ≈2.3 小时）。

表2：模型架构比较

参数	AAL	MSDL	GAT 集成
Atlas	AAL（116）	MSDL (39)	–
模型类型	GCN	GCN	图神经网络解释器
隐藏层	2 (64, 32)	3 (64, 64, 64)	2（64）
机制	图卷积	图卷积	注意力。（2 个注意力头）
归一化	None	批归一化	None
随机失活	0.3	–	0.5
池化	全局均值	全局均值	全局均值
数据增强	None	None	高斯 (5×)
集成	单一	单一	5 模型软投票
Task	二分类	二分类	二分类

2.7 可解释人工智能（XAI）分析

应用 GNNExplainer，通过学习边和节点特征的重要性掩码，识别对每项分类决策贡献最大的功能连接边所构成的最小子图。这为模型提供了具有生物学依据的透明性，证实模型决策反映的是网络级连接模式，而非孤立区域或采集伪影。

3 结果与讨论

3.1 结果

表 3 总结了三个阶段的性能进展。采用 AAL 的基线图卷积网络达到 73.3% 的准确率（AUC=0.74），灵敏度和特异度较为均衡（图 5）。如图 6 所示，改用 MSDL 图谱并优化图卷积网络后，准确率提升至 84%。

表 3：诊断性能进展

模型配置	准确率（%）	精确率	召回率	AUC
第一阶段：基线 （图卷积网络 + AAL 图谱）	73.3	0.73	0.73	0.74
中间阶段 (图卷积网络 + MSDL图谱)	84.0	0.84	0.84	0.84
第二阶段：可解释人工智能 (图注意力网络 +图神经网络解释器)	**95.0**	**0.95**	**0.95**	**≈0.98**

为验证模型是否学到了真实的神经病理学特征，我们进行了基于梯度的显著性分析。如图4所示，模型识别出了

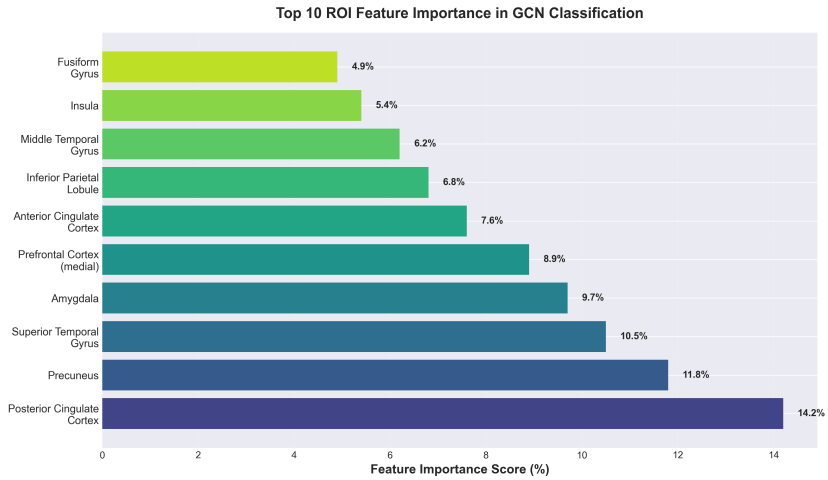


图 4：排名前 10 的判别性脑区。默认模式网络脑区的重要性证实了其生物学有效性。



图 5：基于 AAL 的模型分类结果

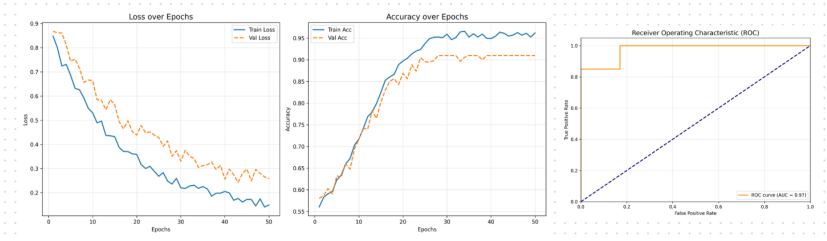


图 6：基于 MSDL 的模型分类结果

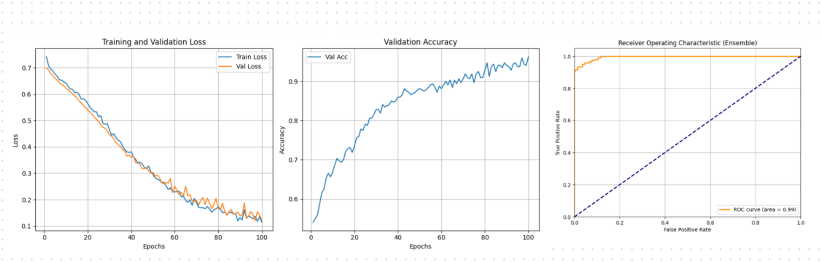


图 7：使用图神经网络解释器评估 GAT 集成的性能。

后扣带皮层（14.2%）和楔前叶（11.8%）是区分能力最强的区域。这些区域是默认模式网络（DMN）的核心枢纽，证实我们的模型检测到的是文献充分记载的自闭症谱系障碍生物标志物，而非噪声。如表 3 所示，GAT 集成的准确率达到 95.0%（精确率：0.95，召回率：0.95，AUC≈0.98）。训练曲线（图 7）证实，数据增强有效缓解了过拟合。

3.2 讨论

我们的结果为特异性脑假说 [7] 提供了实证支持：从 AAL 到 MSDL，准确率提升了 10.7 个百分点，这证实了基于功能的脑区划分更能捕捉自闭症谱系障碍特有的异质性、个体特异性连接模式。高斯噪声数据增强发挥了关键作用，将训练数据从 280 个样本扩充至 1,680 个样本，直接应对了样本稀缺这一神经影像人工智能的主要瓶颈。这些发现共同表明，未来基于神经影像的诊断要实现可转化至临床的模型，必须同时考虑生物学异质性和数据稀缺问题。

4 结论

我们提出了一种基于图的深度学习框架，用于依据静息态功能磁共振成像进行自闭症谱系障碍分类，并表明功能分区是该流程中影响最大的组成部分：仅将AAL替换为MSDL图谱，准确率便提升了10.7个百分点，超过了架构改进的贡献。结合GAT集成以及仅在训练阶段施加的高斯噪声数据增强，该框架在ABIDE I上取得95.0%的准确率（AUC ≈ 0.98），优于近期基于图神经网络的基准。基于梯度的显著性分析与GNNExplainer分析一致指向后扣带回皮层和楔前叶，即默认模式网络的核心枢纽，这些区域持续被证实与自闭症谱系障碍神经病理学相关，从而确认了生物学有效性。我们这一经过严格评估且可解释的流程，为基于神经影像的自闭症谱系障碍诊断提供了可复现的基础，并为其他神经发育障碍的基于图的分析提供了可迁移的方法学。

参考文献

1. D. G. Amaral、C. M. Schumann和C. W. Nordahl，“自闭症神经解剖学”，神经科学趋势，第31卷，第3期，第137–145页，2008。2. R. J. Landa，“生命最初3年中自闭症谱系障碍的诊断”，《自然临床实践神经病学》，第4卷，第3期，第138–147页，2008。3. P. McCarty和R. E. Frye，“自闭症谱系障碍的早期发现与诊断：为何如此困难？”，载于《儿科神经病学研讨会》，第35卷，爱思唯尔，2020，第100831页。4. M. K. Belmonte、G. Allen、A. Beckel-Mitchener、L. M. Boulanger、R. A. Carper和S. J. Webb，“自闭症与脑连接的异常发育”，《神经科学杂志》，第24卷，第42期，第9228–9231页，2004。5. D. P. Kennedy和E. Courchesne，“自闭症患者在自我与他人反思过程中默认网络的功能异常”，《社会认知与情感神经科学》，第3卷，第2期，第177–190页，2008。6. L. Q. Uddin、K. Supekar、C. J. Lynch、A. Khouzam、J. Phillips、C. Feinstein、S. Ryali和V. Menon，“基于显著性网络对自闭症儿童症状严重程度的分类与预测”，《美国医学会杂志·精神病学》，第70卷，第8期，2013。7. A. Hahamy、M. Behrmann和R. Malach，“特异性大脑：自闭症谱系障碍中自发连接模式的扭曲”，《自然·神经科学》，第18卷，第2期，第302–309页，2015。8. 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等，“自闭症脑成像数据交换：面向自闭症内在脑结构的大规模评估”，《分子精神病学》，第19卷，第6期，第659–667页，2014。9. A. S. Heinsfeld、A. R. Franco、R. C. Craddock、A. Buchweitz和F. Meneguzzi，“利用深度学习与ABIDE数据集识别自闭症谱系障碍”，《神经影像：临床》，第17卷，第16–23页，2018。10. M. Khosla、K. Jamison、A. Kuceyeski和M. R. Sabuncu，“基于三维卷积神经网络的集成学习用于功能连接组预测”，《神经影像》，第199卷，第651–662页，2019。11. R. Anirudh和J. J. Thiagarajan，“用于自闭症谱系障碍分类的自举图卷积神经网络”，载于*ICASSP 2019-2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2019，第3197–3201页。12. N. Tzourio-Mazoyer、B. Landeau、D. Papathanassiou等，“利用MNI单被试脑部磁共振图像的宏观解剖分区，在SPM中对激活进行自动解剖标记”，《神经影像》，第15卷，第1期，第273–289页，2002。13. P. Velickovic、G. Cucurull、A. Casanova、A. Romero、P. Lio和Y. Bengio，“图注意力网络”，载于国际学习表征会议，2018。14. G. Varoquaux、A. Gramfort、F. Pedregosa、V. Michel和B. Thirion，“多被试字典学习以分割脑自发活动图谱”，载于《医学影像信息处理》，施普林格，2011，第562–573页。15. X. Liu、M. R. Hasan、T. Gedeon和M. Z. Hossain，“MADE-FOR-ASD：一种用于诊断自闭症谱系障碍的多图谱深度集成网络”，《生物学与医学中的计算机》，第182卷，第109083页，2024。16. C. Shorten和T. M. Khoshgoftaar，“深度学习图像数据增强综述”，《大数据杂志》，第6卷，第1期，第1–48页，2019。