IER-GCN: INVARIANT EDGE RATIONALE FOR ROBUST POPULATION GRAPHS IN MULTISITE NEUROIMAGING

Tianshu Chu[1] *Youyong Kong*[2]

[1] Department of Computer Science, Columbia University in the City of New York
[2] School of Computer Science and Engineering, Southeast University
tc3396@columbia.edu, kongyouyong@seu.edu.cn

ABSTRACT

Multisite neuroimaging cohorts such as ABIDE exhibit substantial OOD (site-level domain) shifts, which can lead graph neural networks to exploit shortcut edges and degrade cross-site generalization. We present IER-GCN, a plug-in framework that integrates Edge-Variational GCNs (EV_GCN) [1] with the Discovering Invariant Rationales (DIR) paradigm [2]. IER-GCN constructs posterior edge scores by combining EV_GCN's pairwise affinity with a residual scorer, transforms them into a differentiable gate that modulates message passing, and trains two heads: a causal head on gated edges and a shortcut head regularized via cross-site interventions drawn from a memory bank. We adopt a strict site-held-out protocol on ABIDE [3] and MDD [4]: train on the train-only subgraph, select on train ∪ val subgraph, and evaluate on the full graph. Experiments show consistent LOSO OOD improvements over EV_GCN and other representative method (LG-GNN, GATE) while yielding stable, interpretable rationales with minimal architectural changes.

Index Terms— Graph neural networks, invariant rationales, ABIDE, MDD, LOSO.

1. INTRODUCTION

Learning from multisite neuroimaging data remains challenging because acquisition differences across scanners, protocols, and demographics induce pronounced out-of-distribution (OOD) shifts. In resting-state fMRI (rs-fMRI), these site-level shifts propagate into population graphs where spurious or site-specific edges can dominate message passing, thereby degrading cross-site generalization. While convolutional networks excel on grid-structured images, population-level disease analysis often benefits from representing subjects as nodes linked by clinically meaningful affinities, which naturally motivates graph neural networks (GNNs). Among GNN-based approaches, Edge-Variational GCNs (EV_GCN) offer a principled way to learn edge weights from non-imaging covariates through a pairwise association encoder and to propagate signals with spectral graph convolutions, providing uncertainty-aware disease prediction in population graphs [1].

However, uncertainty modeling alone does not prevent a model from exploiting shortcut edges that correlate with site rather than diagnosis. Recent work on *invariant rationales* argues that robust predictions should rest on a subset of features or substructures whose predictive role is stable across domains; the DIR paradigm operationalizes this idea by identifying and training on such invariant supports while explicitly countering shortcut signals. Yet, directly transplanting DIR into edge-probabilistic GNNs faces two obstacles: (i) rationale selection must be compatible with continuous, learned edge weights; and (ii) the training protocol must avoid structural leakage from test sites, which is easy to introduce when graph construction and normalization mix domains.

This paper introduces IER-GCN, a plug-in framework that fuses the architectural advantages of EV_GCN with the *full* DIR rationale pipeline. IER-GCN constructs posterior edge scores by combining EV_GCN's learned pairwise affinities with a lightweight residual scorer on pair features; these posteriors are turned into a soft edge mask that rescales PAE weights before filtering without rewriting the backbone. To separate causal and shortcut signals, we train a causal head on gated graphs and a shortcut head on S-dominant graphs, then inject site-swapped subgraphs from a memory bank to quantify sensitivity. Crucially, we enforce a strict leave-one-site-out (LOSO) protocol: training on the train-only subgraph (including normalization), early selection on the train ∪ validation subgraph, and final evaluation on the full graph.

Our main contributions are:

- **Posterior edge gating with dual-head training.** We estimate edge posteriors by combining EV_GCN affinities with a residual scorer, inject a differentiable gate that reweights edges before filtering, and couple a causal head with a shortcut head regularized by cross-site interventions.
- **Leakage-free LOSO pipeline and graph hygiene.** We separate train/validation/test graphs, fit normalization on train-only edges, remove self-loops consistently, and disable internal edge dropout in DIR to avoid shape mismatches and structural leakage.
- **ABIDE and MDD evaluation under OOD.** On ABIDE and MDD datasets, IER-GCN improves LOSO performance over EV_GCN baselines while producing stable, interpretable edge rationales under site-level OOD shift.

2. RELATED WORK

Population graphs for neuroimaging. Population-based disease prediction models represent each subject as a node and integrate imaging with non-imaging covariates to define edges, enabling message passing across clinically related individuals. Early frameworks established this paradigm with spectral GCNs on ABIDE and ADNI [5]. Building on this direction, Edge-Variational GCNs (EV_GCN) learn pairwise affinities via a pairwise association encoder and propagate with spectral convolutions, offering uncertainty-aware prediction by modeling edge variability [1].

Rationales and invariant reasoning for GNNs. Post-hoc explanation methods such as GNNExplainer identify compact subgraphs and features that drive a model's prediction [6]. In contrast, the DIR paradigm seeks *invariant rationales* by constructing interventional distributions and encouraging predictors to rely on substructures that remain stable across environments, thereby improving interpretability and out-of-distribution generalization [2]. Our work adapts DIR to edge-probabilistic GNNs by introducing differentiable posterior gating that interfaces cleanly with EV_GCN.

OOD under multisite settings. In multisite rs-fMRI, scanner/protocol/demographic shifts produce domain-level OOD differences that challenge generalization. By combining posterior edge gating with cross-site interventions, our approach targets edges that transfer across sites while suppressing site-specific shortcuts, complementing prior population-graph and rationale-based methods [5, 1, 2, 6].

Broader brain-graph modeling. *BrainNetCNN* introduced connectome-specific convolutions (edge–edge, edge–node, node–graph), demonstrating the value of operators tailored to brain-network structure [7]. Subsequent clinical studies leveraged spatio temporal graph learning; for example, Kong *et al.* proposed a spatio temporal GCN on dynamic functional connectivity for MDD diagnosis and treatment-response prediction, underscoring the utility of temporally aware graph representations [8]. Complementarily, multi-site harmonization work has shown that site effects materially influence downstream discrimination on ABIDE, reinforcing the need for methods that mitigate domain shift at the graph level [9].

3. METHOD: IER-GCN

3.1. Problem Setup and Graph Construction

We consider a population graph $G = (V, E)$, where each node $i \in V$ represents a subject with imaging feature vector $\mathbf{x}_i$ (e.g., ROI-level descriptors), and each unordered pair $(i, j) \in E$ is annotated by non-imaging pairwise features $\mathbf{z}_{ij}$ (e.g., covariate differences). The goal is binary diagnosis prediction, $y_i \in \{0, 1\}$, under a leave-one-site-out (LOSO) OOD protocol: we train on sites S_{train} and evaluate on an unseen site S_{test}.

Self-loops. We remove self-loops (i, i) prior to any edge computation and normalization. This choice is principled for two reasons. First, in population graphs we model *inter-subject* relations; self-connections add no relational evidence and primarily rescale a node's own signal. Second, our backbone uses *Chebyshev Graph Convolutional Networks (GCN)*, where filtering is expressed as a truncated polynomial of the (scaled) Laplacian $\tilde{L}$. The Chebyshev basis includes the identity term $T_0(\tilde{L}) = I$ by construction, so explicit self-loops are unnecessary and can distort normalization factors; popular implementations (e.g., [10]) also remove them internally during layer preparation. Eliminating self-loops *before* computing edge weights and gates thus aligns the theory with practice and prevents length mismatches when overriding edge weights later.

3.2. Pairwise Association Estimation (PAE): Edge Prior

Following EV_GCN, we derive a data-driven prior weight $w_{ij}^{\text{PAE}} \in (0, 1)$ from $\mathbf{z}_{ij}$ via a pairwise encoder $\phi(\cdot)$ [1]. We map $\mathbf{z}_{ij}$ and $\mathbf{z}_{ji}$ into a shared d-dimensional latent (we use d=128), and form a symmetric affinity by cosine similarity:

$$w_{ij}^{\text{PAE}} = \frac{\langle \phi(\mathbf{z}_{ij}),\, \phi(\mathbf{z}_{ji}) \rangle}{\|\phi(\mathbf{z}_{ij})\|\, \|\phi(\mathbf{z}_{ji})\|} \in (0, 1). \tag{1}$$

Intuitively, w_{ij}^{PAE} quantifies non-imaging coherence between subjects i and j and acts as a *soft adjacency* used by the GNN. To avoid leakage, any feature normalization (for $\mathbf{z}_{ij}$ or w^{PAE}) is fit on *train edges only* and then reused for validation/test.

3.3. Posterior Edge Scoring and Differentiable Gating

A high prior affinity does not guarantee *invariance* across sites: some edges may be highly predictive on S_{train} yet behave as shortcuts on S_{test}. We therefore augment the prior with a residual scorer $\psi_{ij} \in \mathbb{R}$ (a lightweight MLP on $\mathbf{z}_{ij}$) and combine them in the *log-odds* domain:

$$\pi_{ij} = \sigma\Big(\operatorname{logit}\left(w_{ij}^{\text{PAE}}\right)/\tau \;+\; \alpha\, \psi_{ij} \Big), \tag{2}$$

where σ is the sigmoid, $\tau > 0$ is a temperature that calibrates the prior's confidence, and $\alpha \geq 0$ balances the residual contribution. Eq. (2) can be interpreted as follows: $\operatorname{logit}(w_{ij}^{\text{PAE}})/\tau$ encodes a (scaled) prior belief, and $\alpha\, \psi_{ij}$ adds data-adaptive *residual evidence*. Thus, higher π_{ij} indicates that edge (i, j) is both prior-supported and residual-consistent across environments.

We then convert the posterior logits to a *differentiable gate* $g_{ij} \in [0, 1]$ using a temperature-controlled sigmoid centered at a detached top-k threshold (smoothly approximating hard selection). The gate modulates message passing by an *edge-weight override*:

$$\tilde{w}_{ij} \;=\; g_{ij} \cdot w_{ij}^{\text{PAE}}, \tag{3}$$

leaving the backbone unchanged while allowing gradients to flow from the prediction loss to the residual scorer through g_{ij}. For model selection and evaluation, we use a *hard* top-k mask derived from π; optionally, we keep a small floor ε on non-selected edges to stabilize the downstream interventions.

3.4. GCN Backbone Architecture

Our backbone is a stack of L *Chebyshev GCN* layers [10]. A single layer of order K applies a polynomial filter of the (scaled) Laplacian $\tilde{L}$ to the hidden representation $\mathbf{H}^{(\ell)} \in \mathbb{R}^{N \times F_\ell}$:

$$\mathbf{H}^{(\ell+1)} = \sum_{k=0}^{K} \theta_k\, T_k(\tilde{L})\, \mathbf{H}^{(\ell)}, \tag{4}$$

$$T_0(\tilde{L}) = I,\; T_1(\tilde{L}) = \tilde{L},\; T_k(\tilde{L}) = 2\tilde{L}\, T_{k-1}(\tilde{L}) - T_{k-2}(\tilde{L}), \tag{5}$$

with learnable coefficients $\{\theta_k\}_{k=0}^{K}$. In practice we form $\tilde{L}$ from the degree-normalized adjacency built with the *gated* weights $\tilde{w}_{ij}$ in (3). Stacking such layers (the "Chebyshev GCN") yields multi-hop receptive fields while maintaining locality. Compared to first-order GCNs that typically *add* self-loops to recover an identity term [11], the Chebyshev basis explicitly includes I via $T_0(\tilde{L})$; hence our earlier decision to remove self-loops is both natural and numerically stable.

3.5. Causal and Shortcut Heads with Cross-Site Interventions

To separate invariant from shortcut structure, we instantiate two prediction heads that share the backbone design but are optimized with different edge emphases:

- **Causal head (C).** Operates on the full graph with *gated* weights $\tilde{w}$ and is trained on training nodes by cross-entropy, $\mathcal{L}_{\text{C}} = \ell(\hat{\mathbf{y}}^{\text{C}}, \mathbf{y})$.

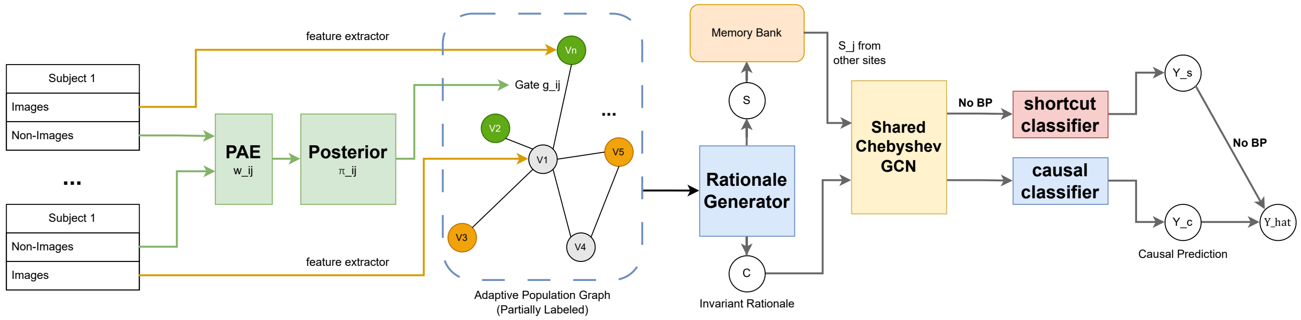

Fig. 1. IER-GCN pipeline

- **Shortcut head (S).** Focuses on the complement (edges with small g_{ij}), forming S-dominant subgraphs. We maintain a memory bank of such S-graphs generated from different S_{train} and perform *cross-site interventions* by replacing S with S_j sampled from other sites.

Given interventions $\{S_j\}_{j=1}^{J}$, we define the training objective

$$\mathcal{L} \;=\; \underbrace{\mathcal{L}_{\mathrm{C}}}_{\text{updates C+residual}} \;+\; \lambda\,\underbrace{\mathrm{Var}_j\big[\mathcal{L}_{S_j}\big]}_{\text{robustness cue}}, \tag{6}$$

$$\mathcal{L}_{S_j} = \ell\big(\hat{\mathbf{y}}^{\mathrm{S}}(S_j), \mathbf{y}\big), \tag{7}$$

with $\lambda \geq 0$. The variance term quantifies the *sensitivity* of predictions to shortcut substitutions; minimizing it encourages reliance on edges whose predictive role is stable across sites, resonating with the DIR principle of invariant rationales [2]. To ensure stable optimization, we compute $\mathrm{Var}_j[\mathcal{L}_{S_j}]$ with gradients stopped for the S branch (used as a scalar signal in the C update), and optimize the S head separately with the mean shortcut loss $\frac{1}{J}\sum_j \mathcal{L}_{S_j}$ using a second optimizer. This decouples gradient flows and avoids double backpropagation through shared computations.

3.6. Training/Validation/Evaluation under LOSO

We construct three edge-disjoint graphs per fold: a *train-only* subgraph for fitting the PAE prior and all normalizations, a *train* ∪ *validation* subgraph for checkpoint selection (hard top-k gate), and the *full* graph for testing. During evaluation, a global top-k can starve test nodes; we therefore enforce a minimum degree for test nodes by promoting their highest-scoring incident edges until a target degree is reached. Throughout DIR training we disable internal edge dropout to keep $|E|$ synchronized with the overridden weights $\tilde{w}$, and we consistently remove self-loops as discussed above.

4. EXPERIMENTS

4.1. Dataset and Protocol

We evaluate on the Autism Brain Imaging Data Exchange (ABIDE), a multisite resting-state fMRI (rs-fMRI) consortium that aggregates data from numerous acquisition centers with heterogeneous scanners and protocols [3]. To obtain subject-level connectomes, we use the publicly released ABIDE Preprocessed resource (CPAC stream; band-pass filtered; nuisance regression without global signal), and extract regional time series with the CC200 atlas; subject-wise ROI–ROI correlations are Fisher z-transformed to form imaging features [12, 13]. Non-imaging covariates (e.g., site, age, sex, and motion summaries when available) serve as pairwise inputs $\mathbf{z}_{ij}$ in the PAE module. Site identifiers follow ABIDE naming, with merged aliases for split collections (e.g., *CMU_a/b* → *CMU*).

We treat each subject as a node; edges connect subject pairs with pairwise covariates $\mathbf{z}_{ij}$. We remove self-loops and compute the PAE prior w_{ij}^{PAE} from $\mathbf{z}_{ij}$, as in Eq. (1). All feature normalization tied to PAE is fit on *train edges only* and reused for validation/test.

We adopt *leave-one-site-out* folds: for each site s, we train on $\bigcup_{t \neq s} \mathcal{D}_t$, select checkpoints on a held-out validation split from the training sites (train ∪ val subgraph), and evaluate on $\mathcal{D}_s$ using the *full* graph. During evaluation we apply hard top-k gating derived from the learned posteriors and enforce a minimum degree for test nodes to avoid isolation after sparsification. Internal edge dropout is disabled in DIR mode to keep the edge set aligned with the overridden weights. Memory-bank interventions for the shortcut head are sampled strictly from training sites.

The backbone is a Chebyshev GCN stack [10]; the gate multiplies learned PAE weights prior to graph filtering, and a lightweight classifier produces node-wise logits. Hyperparameters (e.g., polynomial order K, number of layers L, temperatures τ, τ_g, fusion weight α, sparsity/top-k ratio, and variance coefficient λ) are tuned on the validation split, and early stopping is based on validation accuracy.

We report standard classification metrics on the test site: accuracy (Acc), area-under-ROC (AUC), sensitivity (Sens), specificity (Spec), and F1-score. Let TP, TN, FP, FN denote counts on the test set. Then

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},$$

$$\mathrm{Sens} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{F1} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}.$$

AUC is computed from the ROC induced by the predicted probabilities. These performances are summarized by the mean (and standard deviation) across LOSO folds.

We further evaluate on an rs-fMRI major depressive disorder (MDD) dataset using the same LOSO protocol and the same metrics.

4.2. Results and Analysis

We compare the EV_GCN backbone (PAE prior + Chebyshev GCN; no rationale gating) against our proposed IER-GCN (posterior edge scoring, soft gate during training and hard top-k at validation/test, variance regularization), and additionally include LG-GNN [14] and GATE [15] as representative graph neural network baselines.

To reduce variance, for ABIDE we repeat each method across **five** independent runs and report the mean ± standard deviation over the **17** LOSO folds.

Method	Acc (%)	AUC (%)	Sens (%)	Spec (%)	F1 (%)
LG-GNN [14]	68.98 ± 1.68	67.95 ± 1.31	61.66 ± 2.49	74.25 ± 0.92	68.63 ± 1.41
GATE [15]	63.90 ± 1.28	65.11 ± 2.80	61.66 ± 1.67	64.51 ± 2.29	62.32 ± 1.76
EV_GCN (baseline)	66.40 ± 0.47	68.35 ± 0.69	68.18 ± 1.06	66.79 ± 1.51	66.55 ± 1.24
IER-GCN (ours)	**69.78** ± **1.46**	**69.78** ± **1.38**	68.23 ± 1.34	72.58 ± 2.18	**69.64** ± **1.47**

Table 1. LOSO OOD performance on ABIDE (mean ± std over 17 sites × 5 seeds).

We observe that IER-GCN consistently improves **accuracy** by $+\mathbf{3.38}$ points and **AUC** by $+\mathbf{1.43}$ points over EV_GCN. The most marked change is in **specificity** ($+5.79$ points; 72.58% vs. 66.79%), whereas **sensitivity** is effectively preserved (68.23% vs. 68.18%). This pattern indicates fewer false positives without sacrificing true-positive detection, and is consistent with our design: gating down site-linked edges and stabilizing predictions via cross-site interventions. The **F1**-score improves by $+\mathbf{3.09}$ points, reflecting a better precision–recall balance. Variability across seeds remains modest (e.g., Acc std 1.46%), suggesting stable optimization.

Compared to LG-GNN and GATE, IER-GCN achieves the best Acc/AUC/F1, while LG-GNN attains slightly higher specificity.

On a representative seed, IER-GCN improves AUC on **13/17** held-out sites (median $+4.97$ points). The largest gains are observed for *Caltech* ($+16.0$), *CMU* ($+13.3$), *SBL* ($+13.1$), *UCLA* ($+11.4$), *Olin* ($+10.2$), and *NYU* ($+9.4$). These are folds with pronounced domain shift (scanner/protocol or cohort composition) where suppressing site-linked edges appears particularly beneficial. Performance drops appear at *MaxMun* (-23.6), *Yale* (-11.9), *USM* (-9.8), and mildly at *Stanford*: -0.6. We hypothesize that a global hard top-k might oversparsify graphs at smaller or idiosyncratic sites, and that residual site bias in the shortcut bank may persist.

Variant	Acc (%)	AUC (%)	Sens (%)	Spec (%)	F1 (%)
IER-GCN (best)	**71.76**	**71.82**	**70.23**	**73.02**	**70.99**
No residual (posterior=prior)	69.69 (−2.07)	70.51 (−1.31)	68.18 (−2.05)	72.81 (−0.21)	69.58 (−1.41)
No variance (λ=0)	69.35 (−2.41)	67.23 (−4.59)	66.62 (−3.61)	70.27 (−2.75)	67.58 (−3.41)
No min-degree at test	68.77 (−2.99)	68.16 (−3.66)	67.27 (−2.96)	70.68 (−2.34)	68.36 (−2.63)

Table 2. Ablation study for this work. This table shows how different factors affect the performance of our method. Δ indicates the change vs. IER-GCN (best).

We further examined IER-GCN by ablating three components and comparing each to the best-performing IER-GCN run (Acc 71.76, AUC 71.82, Sens 70.23, Spec 73.02, F1 70.99). (i) *No residual* (posterior = prior) yields Acc 69.69 and AUC 70.51, i.e., $-\mathbf{2.07}$ and $-\mathbf{1.32}$ points versus IER-GCN, with Sens -2.05 and F1 -1.41 points (Spec decreases slightly: -0.21). This indicates that, although the PAE prior is strong on ABIDE, the residual scorer contributes complementary evidence that improves ranking quality and the precision–recall trade-off. (ii) *No variance* (λ=0) shows the largest degradation in AUC and F1 (67.23 and 67.58; $-\mathbf{4.59}$ and $-\mathbf{3.41}$ points), accompanied by drops in Acc (69.35; -2.41), Sens (66.62; -3.61) and Spec (70.27; -2.75). This aligns with our design: removing the variance-based robustness cue weakens invariance pressure under LOSO OOD, allowing shortcut edges to re-enter the decision path and reducing cross-site ranking stability. (iii) *No min-degree at test* reduces AUC to 68.16 and Acc to 68.77 ($-\mathbf{3.66}$ and $-\mathbf{2.98}$ points), with Sens/Spec/F1 also down ($-2.96/-2.34/-2.63$ points). This confirms the practical role of a structural safeguard at evaluation: after hard top-k gating, enforcing a minimum test-node degree prevents isolation on small or heterogeneous sites and preserves discriminative connectivity. Taken together, these ablations substantiate the necessity of our *full* IER-GCN design: the posterior gate leverages residual evidence beyond PAE, the variance term provides a targeted OOD regularizer that most strongly benefits AUC/F1, and the minimum-degree constraint stabilizes performance when the learned rationale is sparsified at test time.

Method	Acc (%)	AUC (%)	Sens (%)	Spec (%)	F1 (%)
LG-GNN [14]	65.24 ± 1.57	65.40 ± 1.61	64.19 ± 2.18	66.02 ± 1.43	64.99 ± 1.75
GATE [15]	60.36 ± 1.92	60.69 ± 1.75	59.27 ± 2.32	61.43 ± 1.86	60.08 ± 2.08
EV_GCN (baseline)	66.50 ± 0.63	67.14 ± 0.84	66.40 ± 1.96	66.90 ± 1.55	65.91 ± 1.65
IER-GCN (ours)	**70.95** ± **1.43**	**73.69** ± **1.55**	64.72 ± 1.64	**75.77** ± **1.84**	**69.44** ± **1.58**

Table 3. LOSO OOD performance on MDD (mean ± std over 7 sites × 5 seeds)

On MDD, IER-GCN again achieves the best Acc/AUC/F1, with a notable specificity gain, consistent with suppressing site-linked shortcut edges.

5. CONCLUSION

Multisite rs-fMRI cohorts pose two persistent obstacles for population-graph learning: (i) site-induced OOD shifts that encourage reliance on shortcut edges, and (ii) the risk of structural/data leakage when graph construction, normalization, and evaluation are not cleanly separated. We introduced IER-GCN, a plug-in framework that integrates a data-driven edge prior with invariance-oriented rationales. Concretely, IER-GCN forms *posterior* edge scores by combining the PAE prior with a lightweight residual scorer, converts these into a differentiable gate that rescales edge weights before Chebyshev GCN, and trains a causal head alongside a shortcut head regularized by cross-site interventions and a variance penalty. A leakage-free LOSO protocol—train-only normalization, train ∪ val selection, full-graph testing—and a minimum-degree safeguard at evaluation complete the pipeline.

On ABIDE, IER-GCN consistently improves accuracy and AUC over EV_GCN, with the largest gain in specificity while maintaining sensitivity, indicating fewer false positives without sacrificing true-positive detection. Per-site analyses show benefits on the majority of held-out sites, and ablations confirm the necessity of each component: removing the variance term or the evaluation safeguard erodes AUC and F1, and collapsing the posterior to the prior reduces ranking quality. Together, these findings show that IER-GCN directly targets the central OOD challenge in multisite neuroimaging while imposing minimal architectural overhead.

Future work will investigate adaptive sparsification (node- or site-aware top-k and degree targets), calibration and uncertainty estimation for the edge gate, test-time adaptation to unseen sites, and broader validation on additional cohorts and modalities (e.g., ABIDE-II, HBN, multimodal phenotypes). Extending the rationale paradigm to multi-label settings and exploring self-supervised pretraining for pairwise encoders are also promising directions.

6. REFERENCES

[1] Yongxiang Huang and Albert CS Chung, "Edge-variational graph convolutional networks for uncertainty-aware disease prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 562–572.

[2] Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua, “Discovering invariant rationales for graph neural networks,” *arXiv preprint arXiv:2201.12872*, 2022.

[3] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al., “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism,” *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[4] 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*, vol. 116, no. 18, pp. 9078–9083, 2019.

[5] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrerro Moreno, Ben Glocker, and Daniel Rueckert, “Spectral graph convolutions for population-based disease prediction,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 177–185.

[6] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec, “Gnnexplainer: Generating explanations for graph neural networks,” *Advances in neural information processing systems*, vol. 32, 2019.

[7] 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*, vol. 146, pp. 1038–1049, 2017.

[8] Youyong Kong, Shuwen Gao, Yingying Yue, Zhenhua Hou, Huazhong Shu, Chunming Xie, Zhijun Zhang, and Yonggui Yuan, “Spatio-temporal graph convolutional network for diagnosis and treatment response prediction of major depressive disorder from functional connectivity,” *Human brain mapping*, vol. 42, no. 12, pp. 3922–3933, 2021.

[9] Sara Saponaro, Alessia Giuliano, Roberto Bellotti, Angela Lombardi, Sabina Tangaro, Piernicola Oliva, Sara Calderoni, and Alessandra Retico, “Multi-site harmonization of mri data uncovers machine-learning discrimination capability in barely separable populations: An example from the abide dataset,” *NeuroImage: Clinical*, vol. 35, pp. 103082, 2022.

[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” *Advances in neural information processing systems*, vol. 29, 2016.

[11] TN Kipf, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[12] R Cameron Craddock, G Andrew James, Paul E Holtzheimer III, Xiaoping P Hu, and Helen S Mayberg, “A whole brain fmri atlas generated via spatially constrained spectral clustering,” *Human brain mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[13] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al., “The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives,” *Frontiers in Neuroinformatics*, vol. 7, no. 27, pp. 5, 2013.

[14] 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*, vol. 42, no. 2, pp. 444–455, 2022.

[15] Liang Peng, Nan Wang, Jie Xu, Xiaofeng Zhu, and Xiaoxiao Li, “Gate: Graph cca for temporal self-supervised learning for label-efficient fmri analysis,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 391–402, 2022.