Article

Graph Neural Network Learning on the Pediatric Structural Connectome

Anand Srinivasan [1,†], Rajikha Raja [1,†], John O. Glass [1], Melissa M. Hudson [2], Noah D. Sabin [1], Kevin R. Krull [3] and Wilburn E. Reddick [1,*

1 Departments of Radiology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; anand.srinivasan@yale.edu (A.S.); rajikha.raja@stjude.org (R.R.); john.glass@stjude.org (J.O.G.); noah.sabin@stjude.org (N.D.S.)
2 Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; melissa.hudson@stjude.org
3 Department of Psychology and Behavioral Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; kevin.krull@stjude.org
* Correspondence: gene.reddick@stjude.org; Tel.: +1-901-595-3276
† Joint First Authors: A.S. and R.R. These authors contributed equally to this work.

Abstract: Purpose: Sex classification is a major benchmark of previous work in learning on the structural connectome, a naturally occurring brain graph that has proven useful for studying cognitive function and impairment. While graph neural networks (GNNs), specifically graph convolutional networks (GCNs), have gained popularity lately for their effectiveness in learning on graph data, achieving strong performance in adult sex classification tasks, their application to pediatric populations remains unexplored. We seek to characterize the capacity for GNN models to learn connectomic patterns on pediatric data through an exploration of training techniques and architectural design choices. Methods: Two datasets comprising an adult BRIGHT dataset (N = 147 Hodgkin's lymphoma survivors and N = 162 age similar controls) and a pediatric Human Connectome Project in Development (HCP-D) dataset (N = 135 healthy subjects) were utilized. Two GNN models (GCN simple and GCN residual), a deep neural network (multi-layer perceptron), and two standard machine learning models (random forest and support vector machine) were trained. Architecture exploration experiments were conducted to evaluate the impact of network depth, pooling techniques, and skip connections on the ability of GNN models to capture connectomic patterns. Models were assessed across a range of metrics including accuracy, AUC score, and adversarial robustness. Results: GNNs outperformed other models across both populations. Notably, adult GNN models achieved 85.1% accuracy in sex classification on unseen adult participants, consistent with prior studies. The extension of the adult models to the pediatric dataset and training on the smaller pediatric dataset were sub-optimal in their performance. Using adult data to augment pediatric models, the best GNN achieved comparable accuracy across unseen pediatric (83.0%) and adult (81.3%) participants. Adversarial sensitivity experiments showed that the simple GCN remained the most robust to perturbations, followed by the multi-layer perceptron and the residual GCN. Conclusions: These findings underscore the potential of GNNs in advancing our understanding of sex-specific neurological development and disorders and highlight the importance of data augmentation in overcoming challenges associated with small pediatric datasets. Further, they highlight relevant tradeoffs in the design landscape of connectomic GNNs. For example, while the simpler GNN model tested exhibits marginally worse accuracy and AUC scores in comparison to the more complex residual GNN, it demonstrates a higher degree of adversarial robustness.

1. Introduction

The human structural connectome represents an individual's brain connectivity network by quantifying the white matter tracts connecting anatomical gray matter regions in the brain using diffusion-weighted magnetic resonance imaging (DW-MRI). Structural connectome data has been a valuable tool for linking brain structure to neurological function and dysfunction [1]. Specifically, it has enhanced our understanding of typical brain development [2,3] as well as various neurological disorders, including epilepsy [4–6], schizophrenia [7,8], Alzheimer's disease [9–12], and autism [12,13].

Due to its promising clinical relevance, the structural connectome has been the focus of multiple machine learning studies [14–16]. Among these, structural connectome classification stands out as a clinically important task, as classification algorithms could facilitate early diagnosis of neurological disorders and improve our understanding of relevant clinical characteristics at a neuroanatomical level. Such analysis employs graph theory, where the structural connectome represents a brain graph G = (V, E), in which brain regions make up the set of vertices V, and tract connections between these regions make up the set of weighted edges E (with the number of tracts between two regions determining the relative weight of the corresponding edge).

Although the structural connectome includes many descriptive cerebral features, its inherent graph structure poses challenges for traditional learning methods. Conventional machine learning and deep learning approaches, which are designed for Euclidean data like text and images, struggle with non-Euclidean graph data due to their inability to capture a graph's topological structure [17]. Graph neural networks (GNNs) were developed to address these shortcomings [18]. GNNs are capable of learning and preserving topological patterns, enabling the training of generalizable, high-performance models on graph data [19].

Sex classification has been a major focal point of previous brain connectivity research [20–22]. Differentiating sex in the structural connectome may lead to a better understanding of neurological disorders with sex-specific presentations. For example, studies have shown sex-specific brain connectivity patterns in patients with autism [23], mild cognitive impairment [24], and conduct disorder [25], which could help explain the differing presentations of these disorders across sexes. Furthermore, studies show differing developmental brain connectivity patterns between male and female adolescents [22,26]. Developing connectivity-based sex prediction models for pediatric patients therefore could improve the understanding of brain development in children and aid in identifying risk factors and early diagnosis of certain neurological disorders [20]. Moreover, since early machine learning works on the structural connectome focused on sex and standard demographic information, this information is routinely available in most publicly available datasets. Sex classification also represents an intuitive classification benchmark for assessing how models perform across structural connectome datasets [27].

While GNN models have achieved reasonable success in classifying sex in adult patients [20,27], this work has not been extended to pediatric populations. Pediatric patients exhibit distinct brain connectivity patterns, particularly in developing white matter tracts, which form the edges of the structural connectome [3,28,29]. Consequently, models trained on adult structural connectome data may not be directly applicable to pediatric patients. Moreover, the typically smaller size of pediatric datasets presents additional challenges in training effective pediatric models.

One approach to address this problem of small sample sizes in pediatric datasets is to use few-shot leaning, which learns to make accurate classifications by training on a very small number of labeled cases when suitable training data is scarce. Few-shot learning approaches for GNNs in connectomics are not well studied [19]. Here, as a few-shot learning approach, a small pediatric dataset was enriched with a much larger adult data to test if GNN models were able to leverage this additional connectomic data to learn to make more accurate predictions.

In this study, GNNs and other standard machine learning models were trained on whole brain structural connectome data for sex classification of both pediatric and adult participants. Classification accuracy for the larger adult dataset was compared to previous studies to compare the performance of the learning models [20,27]. Trained adult, pediatric, and enriched adult/pediatric models were tested on the pediatric dataset [20].

After exploring training approaches using two selected GNN architectures and three other deep and machine learning models for comparison, a series of GNN architecture exploration experiments were conducted to determine the impact of pooling and aggregation function, model depth, and skip connections on the ability for GNNs to capture connectomic patterns. Further, adversarial robustness experiments were conducted to assess how models responded to adversarial attacks. These experiments sought to characterize the landscape of design choices and performance tradeoffs for connectomic models.

The primary objectives of this study are to evaluate whether an enriched training approach enables the GNN to generalize across age groups, achieving pediatric classification accuracy comparable to that of adult GNN models tested on adult datasets and to explore the impact of GNN architectural design choices on connectomic learning ability.

2. Materials and Methods

2.1. Datasets

The Human Connectome Project in Development (HCP-D) sought to characterize brain connectivity over the course of typical development. The pediatric HCP-D dataset included 135 healthy participants, ranging from 8 to 20 years of age with a mean age of 16 years. Participants with health conditions that might have impacted typical development or jeopardized their inclusion within the dataset were excluded. 3D T1 weighted images and DW-MRI sequences were acquired with 3T Siemens Prisma scanners in two shells of b = 1500 and 3000 s/mm^2 along 185 gradient directions [3].

The BRIGHT dataset included 147 adult survivors of childhood Hodgkin's lymphoma (HL) and 162 control participants recruited to frequency match survivors across sex, age, and race/ethnicity. All survivors were 18 years of age or older with a mean age of 36 years. All survivors received thoracic radiation during initial treatment and were 5 or more years from diagnosis at the time of data collection. 3D T1 weighted images and DW-MRI sequences were acquired with a 3T scanner in a single shell of b = 700 s/mm^2 along 30 gradient directions. Demographics from both cohorts are reported in Table 1.

Table 1. Summary statistics for adult BRIGHT and pediatric HCP-D datasets.

Dataset	Patient Population	N	% Female
BRIGHT	Adult	309	53.40
HCP-D	Pediatric	135	56.30

2.2. Structural Connectome Processing

The processing pipeline includes multiple workflows implemented in succession. First, raw DW-MRI data were corrected for noise, motion, and Eddy currents [30]. Images were then analyzed using a spherical deconvolution diffusion model to generate fiber

orientation distribution functions (FOD) and probabilistic tractography was performed to attain whole brain tractograms using the MRtrix3 (https://www.mrtrix.org (accessed on 12 Mar 2023)) iFOD2 algorithm [31]. The whole brain streamlines were filtered using the SIFT2 method to extract realistic streamlines [32]. Whole brain structural connectomes were computed for each subject based on weighted streamline density from the whole brain tractograms and based on the parcellations in the HCP-MMP1.0 atlas consisting of 379 cortical and sub-cortical regions as nodes [33].

2.3. Graph Convolutional Network (GCN) Model Definition

Let a dataset $\boldsymbol{D}$ of size N be defined as $\boldsymbol{D} = (\boldsymbol{D_1}, \boldsymbol{D_2}, \ldots, \boldsymbol{D_N})$. The *i*-th patient in dataset $\boldsymbol{D}$ can be described as $\boldsymbol{D_i} = (\boldsymbol{G_i}, \boldsymbol{X_i}, \boldsymbol{y_i})$, where $\boldsymbol{G_i} = (\boldsymbol{V_i}, \boldsymbol{E_i})$ is the patient's structural connectivity graph with its corresponding node feature matrix $\boldsymbol{X_i} \in \mathbb{R}^{|V_i| \times d}$ and class label $\boldsymbol{y_i} = \{\boldsymbol{0}, \boldsymbol{1}\}$. Any patient graph $\boldsymbol{G_i}$ can also be represented as an adjacency matrix $\boldsymbol{A_i} \in \mathbb{R}^{|V_i| \times |V_i|}$. A model function $\boldsymbol{f_\theta} : (\boldsymbol{G_i}, \boldsymbol{X_i}) \to \hat{\boldsymbol{y}}_i$ was defined where $\boldsymbol{\theta}$ is the set of learnable parameters and $\hat{\boldsymbol{y}}_i$ is the model's predicted class label. The objective was to learn a function $\boldsymbol{f}$ which minimizes the classification loss over the dataset $\boldsymbol{D}$.

This study focuses on GNN models, specifically Graph Convolutional Networks (GCNs). GCNs aim to create s-dimensional node embeddings for each node in a graph $\boldsymbol{G_i}$. These individual node embeddings can subsequently be pooled into a single s-dimensional graph embedding vector, which can finally be fed into a downstream classifier for prediction of the output label. GCNs generate node embeddings through an iterative, message passing algorithm. Nodes send messages to their surrounding neighborhoods, and these messages are aggregated and used to update each node's current embedding.

A single layer of a GCN model can be defined as:

$$H_i^l = \sigma\big(A_i^* H_i^{(l-1)} W + b\big) \tag{1}$$

$$H_i^0 = X_i \tag{2}$$

where $\boldsymbol{A_i^*}$ is the normalized adjacency matrix representation of the patient graph $\boldsymbol{G_i}$, which is derived from structured connectome data. $\boldsymbol{A_i^*}$ is a symmetric matrix with dimensions corresponding to the number of nodes in the graph. In this study, we utilized 379 nodes defined by the HCP-MMP1.0 atlas for the whole brain, resulting in $\boldsymbol{A_i^*}$ being a 379 × 379 matrix. $\boldsymbol{H_i^{(l-1)}}$ is the node embedding matrix from the previous iteration, $\boldsymbol{W}$ is the learnable weight matrix, $\boldsymbol{b}$ is the learnable bias term, $\boldsymbol{\sigma}$ is the non-linear activation function. In this study, we use the ReLU (Rectified Linear Unit) and tanh activation functions for σ. $\boldsymbol{H_i^l}$ is the resulting node embedding matrix of the GCN layer at the *i*-th iteration. The node embedding matrix is initialized with the node feature matrix in the first iteration. In the above equation, $\boldsymbol{H_i^{(l-1)} W}$ represents the messages passed from each node to its surrounding neighborhood, and $\boldsymbol{A_i^* H_i^{(l-1)} W}$ represents the weighted aggregation of these outgoing messages based upon the strength of the connection between nodes in the normalized adjacency matrix. Individual GCN layers can be stacked or joined via skip connections to build larger networks.

2.4. Model Architecture

Two GCN models (simple [34] and residual), a deep neural network (multi-layer perceptron (MLP)), and two standard machine learning models (random forest (RF) and support vector machine (SVM)) were trained. Random Forest and SVM were selected as baseline models to enable a meaningful comparison with GNNs. Random Forest was chosen for its resilience to noise, capability to manage high-dimensional data like structural connectome features. SVM was included for its proven effectiveness in binary classification tasks and its capacity to capture non-linear relationships through kernel methods. Unlike

GNNs, which directly utilize the graph structure of the connectome, these models rely on predefined features, offering a contrasting perspective. This comparison allowed us to highlight the benefits of graph-based learning and demonstrate the superior performance of GNNs in this context. Model architectures for both GCN models and the MLP are shown in Figure 1, and model sizes for these models are displayed in Table 2. The simple GCN model contains GCN layers with 64 neurons each, while the residual GCN model contains GCN layers with 32 neurons each. The simple GCN model uses ReLU activation functions between graph convolutional layers and mean global pooling for the generation of a graph embedding vector, while the residual GCN model uses tanh activation and mean aggregation functions between graph convolutional layers to create intermediate graph embedding matrices. The model architecture for the residual GCN is adapted from a model proposed in a brain connectomics benchmarking paper [5]. For the MLP network, hidden layers with 512, 256, and 128 neurons are used with ReLU activation. All models were implemented with PyTorch Geometric.
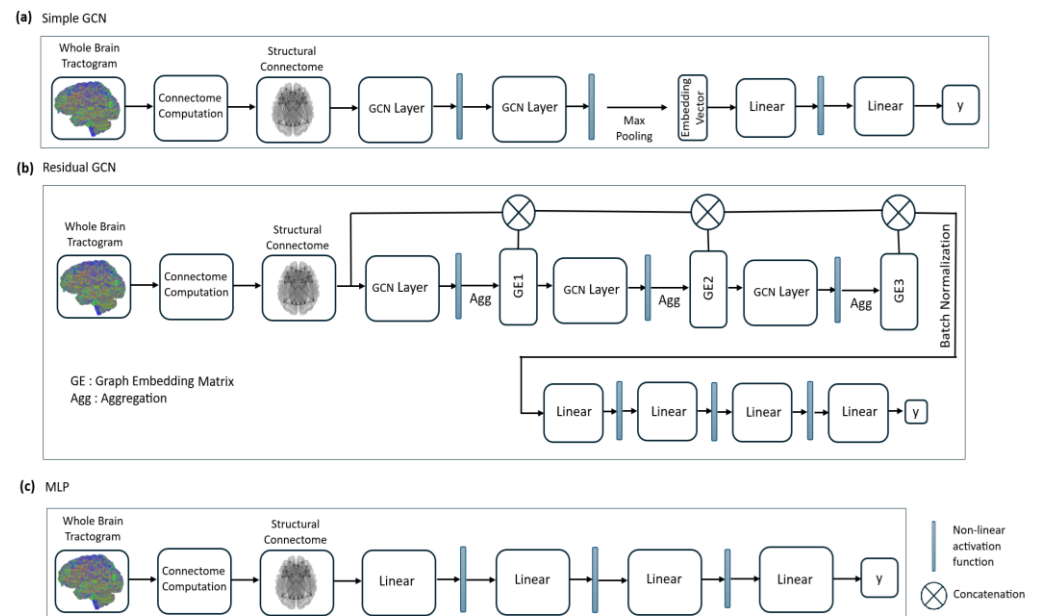

Figure 1. Model architecture illustrations for simple GCN (**a**), residual GCN (**b**), and multi-layer perceptron (**c**) networks.

Table 2. Model sizes for multi-layer perceptron (MLP), simple graph convolutional neural network (GCN simple), and residual graph convolutional neural network (GCN residual).

	MLP	GCN (Simple)	GCN (Residual)
Number of Model Parameters	7.37×10^7	3.89×10^4	4.78×10^6

2.5. Training Procedure

The hyperparameter space for all deep learning models was explored extensively. All deep learning models trained with a learning rate of 1×10^{-3} for 100 epochs using weighted binary cross entropy loss. A number of regularization techniques were employed to mitigate overfitting. 50% dropout was added between GCN layers and fully connected layers for GNN and MLP models, respectively. All deep learning models were trained using a weight decay of 5×10^{-4}, and early stopping was adopted with a patience of 35 epochs.

For the RF and SVM models, principal component analysis was used to reduce connectivity matrix data to 100 dimensions. The random forest model was initialized with

100 estimators, and the support vector machine applies the radial basis function (rbf) kernel. All other parameters are left as default.

Data was split with 70% for training, 10% for validation, and 20% testing in all deep learning experiments. For the two standard machine learning models (RF and SVM), 80% of the data was used for training and the remaining 20% was used for testing. Five-fold cross validation was employed to generate five unique testing sets. The mean accuracy and AUC scores across all sets is reported.

2.6. Model Evaluation Experiments

First, all models were trained and tested against the two available datasets individually. For experiments on individual datasets, data from a single dataset was split into training, validation, and testing sets and each model's mean accuracy and AUC score on the unseen test data is reported. Models trained on the adult BRIGHT dataset were also externally validated on pediatric HCP-D data to test how well structural connectivity models trained solely on adult participants generalize to pediatric participants.

Finally, models were trained and tested on the adult-enriched pediatric dataset. Data from both datasets was shuffled together and split into training, validation, and testing sets. The ratio of adults and pediatric data was held constant across all sets. Each model's accuracy on the unseen test data was reported both as an overall score and stratified by the dataset of origin.

2.7. GNN Architecture Exploration Experiments

After initial model evaluations, a variety of architectural exploration and ablation experiments were conducted to characterize the impact and effect of different model components on performance.

Mean and max pooling and aggregation approaches were compared to investigate the impact of these pooling and aggregation methods on the ability of the GNN models to learn connectomic patterns. The simple GCN, which initially used a mean global pooling function to generate a graph embedding vector, was tested with a max global pooling function replacement, while the residual GCN, which uses a mean node aggregation for graph feature generation, was tested with a max aggregation function replacement. Both tests were performed using the adult-enriched dataset for training and evaluation.

Next, the effect of model depth was investigated for both GNNs. The simple GNN was initially evaluated with two GCN layers. Deeper analogs with additional GCN layers were tested to determine if increasingly depth would improve the model. The residual GCN was initially evaluated with three GCN layers. Two- and four-layer analogs were tested on the adult-enriched dataset to determine the effect of decreasing or increasing the depth on the residual network.

Finally, an ablation experiment was performed to test how skip connections contributed to the ability of the residual GCN model to learn on connectomic data. All skip connections were removed, and the resulting model was trained and tested on the adult-enriched dataset.

2.8. Adversarial Sensitivity Experiments

Adversarial attacks are small, targeted perturbations applied to input data in order to mislead classification models. In the medical domain, robustness to adversarial attacks is important, as errors caused by such attacks could result in harmful or even fatal consequences. Further, a model's sensitivity to small adversarial perturbations may be an indicator of robustness to general noise, which is an inevitable component of real-world medical data.

Adversarial robustness for all deep learning models was evaluated using the fast gradient sign method (FGSM) adversarial attack. FGSM is a white-box adversarial evasion attack which can be defined as:

$$\mathbf{X_i^{adv}} = \mathbf{X}_i + \epsilon * sign(\nabla_X L(\theta, X_i, y_i)) \tag{3}$$

where $\mathbf{X_i}$ is a node feature matrix corresponding to an input graph G_i, L is the loss as function of the model's trainable parameters θ, the node feature matrix X_i and the corresponding label y_i, ϵ is the size of the adversarial perturbation, and $\mathbf{X_i^{adv}}$ is the resulting perturbed node feature matrix.

White box attacks like FGSM enjoy full access to a model's loss function and trainable parameters. As such, they can construct powerful, targeted adversarial examples. Increasing the value of epsilon increases the distance between the original graph and the adversarial example, therefore leading to stronger adversarial attacks. All deep learning models trained on the adult-enriched dataset were tested on target adversarial examples generated by the FGSM method for a range of epsilon values. The average resulting accuracy and AUC scores across the five cross validation test sets are reported.

All code used in this study is publicly available: https://github.com/SrinivasanAnand/GNN_structural_connectivity (accessed on 10 Dec 2024).

3. Results

3.1. Adult Training and Adult Testing

The two standard machine learning models RF and SVM, the deep learning neural network MLP, and the two GCN models (simple and residual) were trained and tested on the adult BRIGHT dataset. The corresponding accuracy and AUC score for each model is shown in Table 3. The GCN models outperform others on the adult dataset, with accuracies exceeding 82%, followed by the multi-layer perceptron model at 77%. GCN models also demonstrate the best AUC scores, exceeding 0.90 for both models.

Table 3. Sex classification mean accuracy and AUC score results for random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), simple graph convolutional neural network (GCN simple), and residual graph convolutional neural network (GCN residual) classifiers trained and tested on the BRIGHT (adult) dataset with errors determined by the sample standard deviation. **Red** indicates overall best results; **bold** indicates overall second-best results.

Metric	RF	SVM	MLP	GCN (Simple)	GCN (Residual)
Accuracy (%)	73.13 ± 2.05	76.37 ± 5.84	77.66 ± 3.18	85.10 ± 2.84	**82.82 ± 5.30**
AUC Score	0.80 ± 0.01	0.86 ± 0.03	0.89 ± 0.01	**0.91 ± 0.03**	0.93 ± 0.03

3.2. Adult Training and Pediatric Testing

All classification models were next trained on the adult BRIGHT dataset and tested on the pediatric HCP-D dataset. Experimental results are shown in Table 4. The residual GCN model remains the most robust, achieving an accuracy above 74% and an AUC score of 0.86. All other models lagged with accuracies of 50–60%. Although the residual GCN exhibits a degree of robustness, all models trained on adult data performed sub-optimally for classification of the pediatric datasets.

Table 4. Sex classification mean accuracy and AUC score results for random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), simple graph convolutional neural network (GCN simple), and residual graph convolutional neural network (GCN residual) classifiers trained on the BRIGHT (adult) dataset and tested on the HCP-D (pediatric) dataset with errors determined by the sample standard deviation. Red indicates overall best results; bold indicates overall second-best results.

Metric	RF	SVM	MLP	GCN (Simple)	GCN (Residual)
Accuracy (%)	55.70 ± 3.23	56.15 ± 3.79	53.93 ± 3.85	**60.74 ± 3.47**	**74.96 ± 2.41**
AUC Score	0.62 ± 0.01	0.64 ± 0.03	0.60 ± 0.01	**0.71 ± 0.03**	**0.86 ± 0.03**

3.3. Pediatric Training and Pediatric Testing

Given the anticipated sub-optimal performance of the adult models for external validation on the pediatric dataset, all classification models were next trained and tested on the pediatric HCP-D dataset. The corresponding accuracy and AUC score for each model is shown in Table 5. The simple GCN model demonstrated the highest accuracy in the low-70% range, while the residual GCN achieved the highest AUC score of 0.85. The residual GCN shows strong discriminative power in terms of AUC score, but all models achieve sub-optimal classification accuracy.

Table 5. Sex classification mean accuracy and AUC score results for random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), simple graph convolutional neural network (GCN simple), and residual graph convolutional neural network (GCN residual) classifiers trained and tested on the HCP-D (pediatric) dataset with errors determined by the sample standard deviation. Red indicates overall best results; bold indicates overall second-best results.

Metric	RF	SVM	MLP	GCN (Simple)	GCN (Residual)
Accuracy (%)	**66.67 ± 8.76**	65.93 ± 8.25	56.30 ± 7.55	**71.11 ± 10.84**	66.67 ± 12.83
AUC Score	0.73 ± 0.13	**0.80 ± 0.04**	0.67 ± 0.05	0.76 ± 0.09	**0.85 ± 0.05**

3.4. Adult-Enriched Pediatric Dataset Training and Testing

Given the sub-optimal performance of the models trained on either the adult or pediatric datasets for classification of the pediatric testing set, an adult-enriched pediatric dataset was then used for both training and testing. Table 6 displays the performance of the models when classifying the HCP-D pediatric data within the adult-enriched pediatric testing dataset. The residual GCN model excels in the pediatric subset of the test set, achieving an accuracy of 83%, and the simple GCN model reaches 79% classification accuracy. These model performances demonstrate a noticeable improvement compared to the next best classifier within this training approach and any other classifier's performance across all training approaches on the HCP-D pediatric dataset.

Table 6. Sex classification mean accuracy and AUC score results for random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), simple graph convolutional neural network (GCN simple), and residual graph convolutional neural network (GCN residual) classifiers trained on the adult-enriched pediatric dataset and tested on the pediatric subset of the test set with errors determined by the sample standard deviation. Red indicates overall best results; bold indicates overall second-best results.

Metric	RF	SVM	MLP	GCN (Simple)	GCN (Residual)
Accuracy (%)	65.93 ± 10.32	67.41 ± 8.25	72.59 ± 6.46	**79.26 ± 6.46**	**82.96 ± 5.02**
AUC Score	0.74 ± 0.01	0.80 ± 0.08	0.79 ± 0.06	**0.84 ± 0.05**	**0.91 ± 0.04**

For completeness, the overall classification accuracies of the models of the adult-enriched pediatric dataset are presented in Figure 2 for the overall testing set and the subsets of adult and pediatrics. Overall, the GCN models demonstrate the strongest performance across accuracy and AUC metrics. Notably, the multi-layer perceptron also exhibits strong classification ability under this training approach.
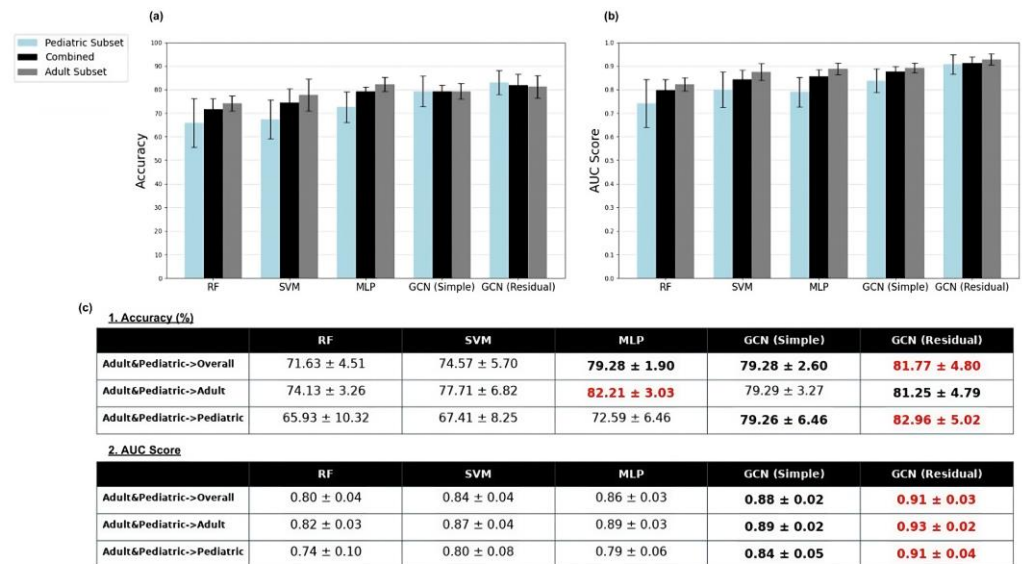

1. Accuracy (%)

	RF	SVM	MLP	GCN (Simple)	GCN (Residual)
Adult&Pediatric->Overall	71.63 ± 4.51	74.57 ± 5.70	79.28 ± 1.90	79.28 ± 2.60	81.77 ± 4.80
Adult&Pediatric->Adult	74.13 ± 3.26	77.71 ± 6.82	82.21 ± 3.03	79.29 ± 3.27	81.25 ± 4.79
Adult&Pediatric->Pediatric	65.93 ± 10.32	67.41 ± 8.25	72.59 ± 6.46	79.26 ± 6.46	82.96 ± 5.02

2. AUC Score

	RF	SVM	MLP	GCN (Simple)	GCN (Residual)
Adult&Pediatric->Overall	0.80 ± 0.04	0.84 ± 0.04	0.86 ± 0.03	0.88 ± 0.02	0.91 ± 0.03
Adult&Pediatric->Adult	0.82 ± 0.03	0.87 ± 0.04	0.89 ± 0.03	0.89 ± 0.02	0.93 ± 0.02
Adult&Pediatric->Pediatric	0.74 ± 0.10	0.80 ± 0.08	0.79 ± 0.06	0.84 ± 0.05	0.91 ± 0.04

Figure 2. Sex classification mean accuracy (**a**) and AUC score (**b**) results for random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), simple graph convolutional neural network (GCN simple), and residual graph convolutional neural network (GCN residual) classifiers trained on the adult-enriched pediatric dataset. Results for pediatric (blue), adult (dark gray), and overall (black) subsets of the test dataset were displayed. Standard deviations of classification accuracy shown by whiskers. AUC and accuracy score results are displayed in tabular format in (**c**). Red indicates overall best results; bold indicates overall second-best results.

Loss curves for all deep learning models from a representative cross validation split for this training approach are shown in Figure 3. All models exhibit an ability to learn connectomic features. Both GCNs show signs of overfitting, but this trend is more pronounced in the residual model.
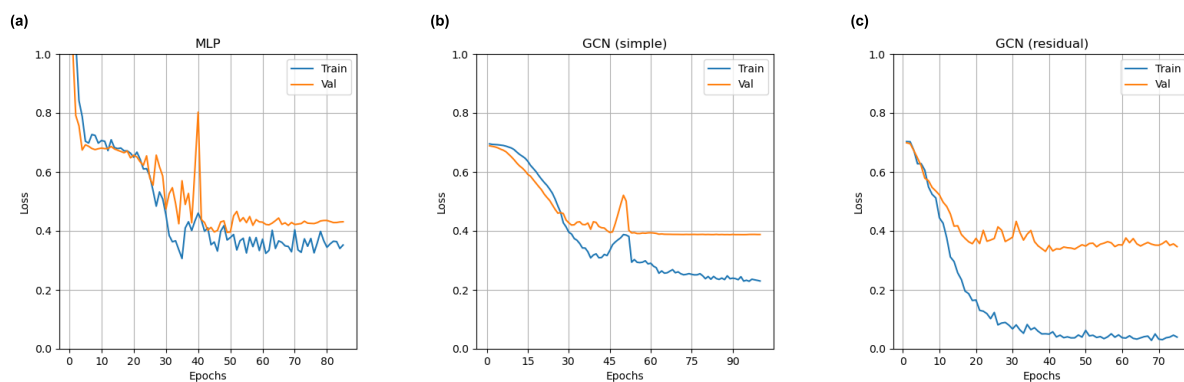

Figure 3. Representative training and validation loss curves for multi-layer perceptron (**a**), simple graph convolutional neural network (**b**), and residual graph convolutional neural network (**c**) classifiers trained on the adult-enriched pediatric dataset.

Receiver operating characteristic (ROC) curves on the training and test set for all models from a representative cross validation split are depicted in Figure 4. Three ROC curves are displayed for each test set, corresponding to the overall test set and the pediatric and adult stratifications. As follows from the AUC score results, the GCN models demonstrate the strongest ROC curves.
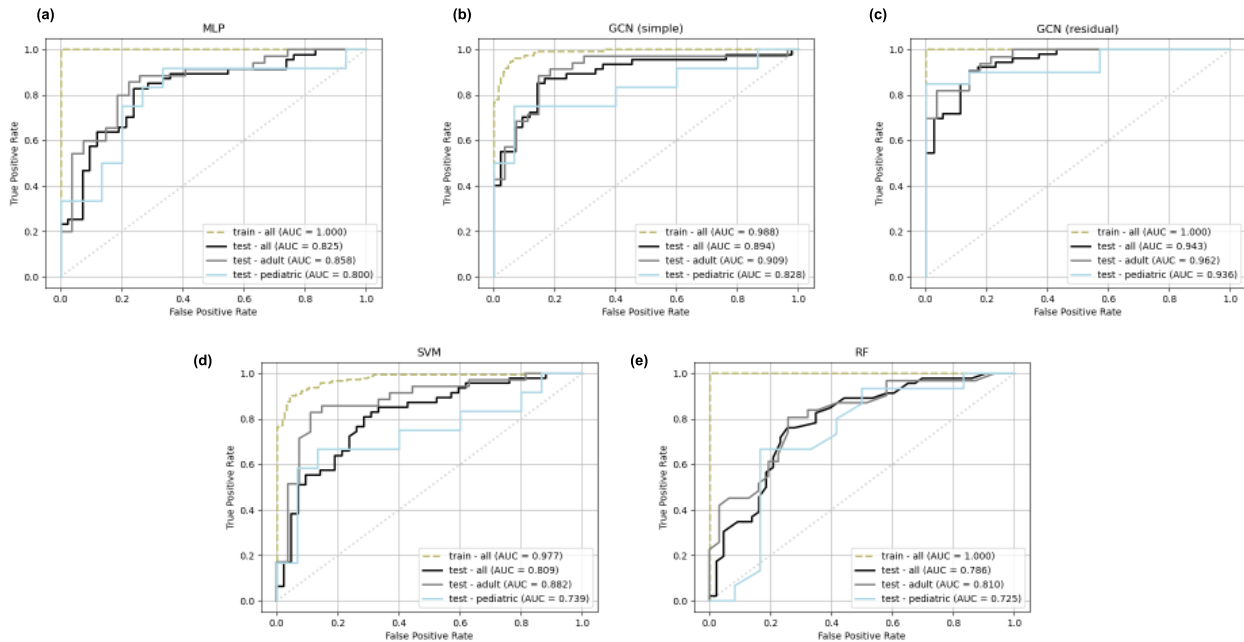

Figure 4. Representative receiver operating characteristic (ROC) curves for multi-layer perceptron (**a**), simple graph convolutional neural network (**b**), residual graph convolutional neural network (**c**), support vector machine (**d**), and random forest (**e**) classifiers trained on the adult-enriched pediatric dataset. Three test set ROC curves are displayed for each model corresponding to pediatric (blue), adult (gray), and overall (black) subsets. A single ROC curve corresponding to the overall training set (green) is displayed for each model.

3.5. GNN Architecture Exploration Results

Architecture exploration experiments were conducted in order to determine the impact of architectural design choices on the ability of the GNN models to learn connectomic patterns. Since the focus of this study is the pediatric connectome, pediatric classification accuracy and AUC scores were used as the primary measure of model performance for exploration experiments. All models were trained on the adult-enriched pediatric dataset since this training method was previously determined to produce the strongest models.

Max and mean pooling and aggregation types were compared across the two GCN models. Tables 7 and 8 display accuracy and AUC scores for the simple and residual GCN models trained with mean and max pooling and aggregation approaches on the combined dataset. The choice of pooling or aggregation type does not show any major effect.

Table 7. Sex classification mean accuracy and AUC score results simple graph convolutional neural network using mean and max pooling functions for graph embedding vector generation.

Pooling Function	Accuracy (%)	AUC Score
Mean	79.26 ± 6.46	0.91 ± 0.05
Max	77.04 ± 7.18	0.85 ± 0.04

Table 8. Sex classification mean accuracy and AUC score results for the residual graph convolutional neural network model using mean and max aggregation functions for graph embedding matrix generation.

Aggregation Function	Accuracy (%)	AUC Score
Mean	82.96 ± 5.02	0.91 ± 0.04
Max	81.48 ± 5.24	0.90 ± 0.04

Simple and residual GCN models of varying depth were tested to assess the impact of depth on learning ability. Results for depth experiments are presented in Figure 5. Again, no major effect is observed, although it is noteworthy that the 3-layer residual network outperforms its 2-layer and 4-layer analogs by 3 accuracy points. For the simple GCN model architecture, shallow networks perform just as well as deeper ones.
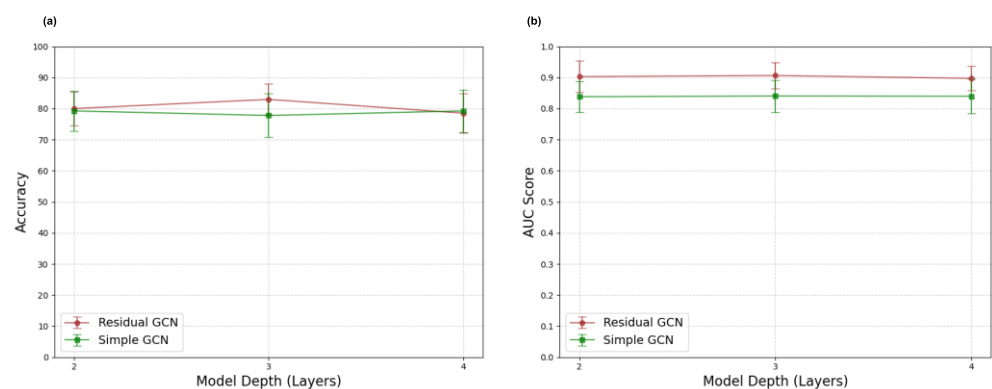

Figure 5. Sex classification mean accuracy (**a**) and AUC score (**b**) results for simple graph convolutional neural network (simple GCN) and residual graph convolutional neural network (residual GCN) architectures with varying model depth analogs.

Skip ablation experimental results are shown in Table 9. Removing skip connections from the residual GCN network results in a large deterioration in performance, with a 14-point accuracy drop on the pediatric stratification of the dataset, highlighting the importance of skip connections for allowing the residual network to learn connectomic patterns.

Table 9. Sex classification mean accuracy and AUC score results for the residual graph convolutional neural network with and without skip connections.

Residual GCN Model (with/without skips)	Accuracy (%)	AUC Score
Without Skips	68.89 ± 9.54	0.74 ± 0.07
With Skips	82.96 ± 5.02	0.91 ± 0.04

3.6. Adversarial Sensitivity Experiment Results

All deep learning models trained on the adult-enriched dataset were tested on adversarial test sets, generated by targeted FGSM attacks on the pediatric stratification of the enriched test dataset. Perturbation sizes between 1.0×10^{-5} and 1.0×10^{-3} were tested. Accuracy and AUC score results for all deep learning models on perturbed examples are shown in Figure 6. The simple GCN model remains the most robust to adversarial attacks, followed by the multi-layer perceptron and the residual GCN, respectively.
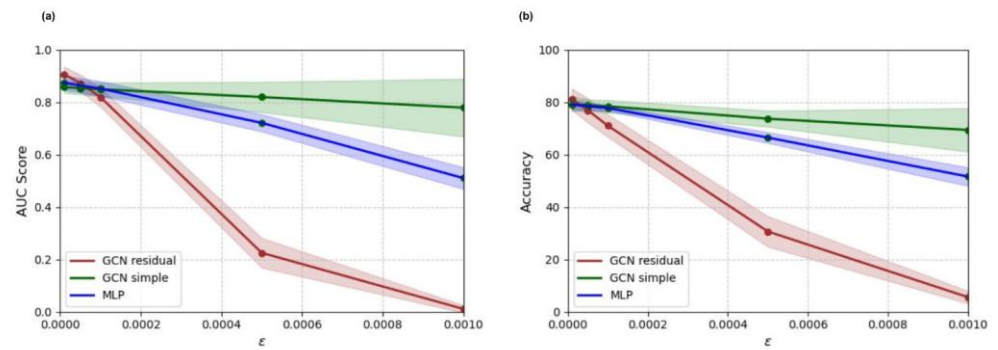
Figure 6. Adversarial accuracy (**a**) and AUC score (**b**) results for multi-layer perceptron (MLP), simple graph convolutional neural network (GCN simple), and residual graph convolutional neural network (GCN residual) classifiers trained on the adult-enriched pediatric dataset.

4. Discussion

Experimental results on the adult BRIGHT dataset indicated that graph neural networks outperform traditional machine learning and deep learning approaches in classifying adult structural connectome data. The highest classification accuracy (85.1%) was achieved by the simple GCN model, which surpasses the best non-GCN model by 7.5%. These findings were consistent with the 86.7% reported in previous studies [20,27].

To assess the robustness of the models trained on the adult dataset, external validation was conducted using the pediatric dataset, allowing us to evaluate the generalizability of these models across different patient age demographics. The results reveal that most models were not robust to this demographic shift, an outcome that was anticipated given the inherent variability between the two datasets. The adult dataset, comprising both HL survivors and community controls, used a slightly different diffusion imaging protocol compared to the pediatric dataset, which included only healthy controls. Moreover, it is well-documented that adults exhibit distinct structural connectivity patterns compared to neurologically developing children and teenagers [3,22,26]. Given these distinctions, it is unsurprising that most models trained on the adult BRIGHT dataset struggled to generalize the HCP-D data.

However, it is noteworthy that the residual GCN model exhibited a degree of robustness in the external validation, maintaining an accuracy of over 74% on the pediatric dataset. This performance suggests that the residual GCN model could capture highly generalizable patterns within graph data, extending applicability from adult participants to their pediatric counterparts. Consequently, robust adult-trained graph neural network connectome models may hold potential for direct application to pediatric populations in future clinical settings. This observed generalizability could be particularly valuable in extending the benefits of future connectome-based algorithms to pediatric participants, especially given that the limited size of pediatric datasets often constrains the development of models trained exclusively on pediatric data.

Experiments on the pediatric HCP-D dataset demonstrated that all models were unable to learn highly generalizable patterns on the pediatric connectome without any form of data enrichment. The best model tested was the simple GCN model, which achieves a classification accuracy of 71.1%. The 14-point performance gap observed between the top-performing adult and pediatric models is likely attributable to the smaller size of the pediatric dataset, a limitation that subsequent experiments aimed to address.

An adult-enriched pediatric dataset training approach was employed to assess whether this method could yield stronger models for pediatric participants. This approach functions as a form of data enrichment, utilizing adult patient data to enhance the

pediatric dataset and thereby facilitate the training of more robust pediatric models. Additionally, a subset of the adult data was reserved to enable testing of each model's performance on both unseen pediatric and adult datasets. The enrichment approach allowed deep learning models to achieve strong classification performance, with the residual GCN reaching 81.8% and the simple GCN and multi-layer perceptron both following at 79.3%. Notably, the residual GCN model also performed exceptionally well across both pediatric and adult participants, attaining an accuracy of 83.0% and 81.3%, respectively. The pediatric test accuracy of the residual GCN model is higher than that of all other pediatric models across all training approaches, narrowing the gap between the best achievable pediatric and adult test accuracies to within about two percentage points (83.0 vs. 85.1%). Thus, enriching smaller pediatric datasets with adult structural connectome data can enhance model performance, enabling the development of pediatric models that achieve performance levels comparable to state-of-the-art adult models. While this study enriches pediatric data using adult data to improve model performance, alternative strategies for addressing the limited availability of pediatric connectome datasets should be explored. One potential approach is to collaborate with other pediatric research institutes or initiatives, which may provide access to larger and more diverse pediatric datasets. Examples include publicly available repositories like the Pediatric Imaging, Neurocognition, and Genetics (PING) study [35] or the ABCD dataset [36], which, although limited in diffusion MRI, could still complement structural connectome analyses. Another promising strategy involves leveraging data augmentation techniques specific to connectome graphs, such as perturbation or simulation of biologically plausible variations. These approaches could mitigate the dependence on adult data enrichment while ensuring that the model remains focused on pediatric specific characteristics.

The limited availability of structural connectome data led to the selection of an adult dataset comprising both HL survivors and community controls, while the pediatric dataset included only healthy controls. These datasets were collected using slightly different DW-MRI protocols including different gradient directions and b-values. Additionally, preprocessing pipelines varied, with adult dataset employing standard motion correction and eddy current correction, while pediatric dataset incorporated advanced susceptibility artifact correction. These variations can impact the derived connectome features, with lower b-values or few-er directions potentially reducing tractography accuracy and voxel resolution differences influencing connectivity matrix representations. Consequently, the observed poor external validation results may reflect these protocol differences rather than a lack of model robustness to the population demographics. While this study focuses on improving model generalizability, future work could benefit from using datasets collected with consistent DW-MRI protocols to better isolate the effects of patient demographics. Alternatively, protocol harmonization techniques, such as advanced preprocessing or domain adaptation methods, could mitigate these challenges and enhance model robustness. Despite these limitations, the inclusion of datasets with differing DW-MRI protocols reflects the practical challenges of applying machine learning models across diverse clinical and research settings. Models that generalize well across such variations are more likely to succeed in real-world applications, making this an important avenue for further investigation. Nevertheless, the residual GCN model's strong external validation performance suggested its capacity to learn highly generalizable patterns that extend from adult to pediatric participants. These generalizable patterns likely represent biologically relevant and stable features within structural connectomes. For instance, the model appeared to leverage connectivity profiles of key hub regions such as the precuneus, posterior cingulate cortex, and superior frontal gyrus, which are central nodes in the brain's structural network and play a crucial role in maintaining global communication [37–39]. These regions are recognized as stable and central nodes, showing relatively low

variability across demographic groups [40,41]. Additionally, strong patterns were reported previously in inter-hemispheric connections, particularly in commissural pathways like the corpus callosum. These connections are highly conserved across individuals and age groups, making them reliable features for model generalization [42–46]. The model likely captured patterns of hierarchical organization, including connectivity differences between primary sensory-motor regions, which mature earlier, and higher-order associative areas, which exhibit prolonged development and greater integration over time. These patterns align with established neurodevelopmental trajectories and demonstrate consistency across age groups [47–51]. These patterns suggest that the residual GCN model leveraged features that are biologically consistent and less variable across demographics, contributing to its generalizability.

The differences between the two datasets may have limited the effectiveness of the adult-enriched pediatric dataset training approach. Ideally, data enrichment should be performed using data that closely resembles the original dataset. The pediatric dataset in this study covered ages 8–20 years, while the adult dataset ranged from 18–65 years with a mean age of 35 years. The overlap between the older pediatric participants (18–20 years) and younger adults (18–20 years) helped bridge the two datasets and reduce discontinuities in connectome patterns. Furthermore, the relatively stable nature of structural connectomes in adulthood supports the use of a broader adult age range [52–54]. To minimize potential biases, stratified sampling was used during training and cross-validation to ensure proportional representation of both groups. While these strategies mitigated bias, future work could refine this approach by exploring harmonization methods or focusing on more age-restricted datasets to ensure even greater neurodevelopmental consistency. Additionally, employing an adult dataset composed solely of healthy controls and collected using the same DW-MRI protocol as the pediatric dataset might have further enhanced the models' performance. Future research should explore the efficacy of the data enrichment approach using other adult and pediatric datasets across a range of tasks to validate its ability to improve pediatric models. Although the enrichment of pediatric datasets using adult structural connectomes improved classification performance, achieving 83% accuracy on pediatric data, further gains might be possible through explicit domain adaptation techniques. Methods such as adversarial domain adaptation, which aligns feature distributions between source (adult) and target (pediatric) domains, could help address inherent differences in brain connectivity patterns between age groups. Additionally, transfer learning approaches, where a model pre-trained on adult data is fine-tuned on pediatric data, could allow the model to learn pediatric-specific features more effectively. These strategies could complement the enrichment approach and enhance the generalizability and robustness of the GNN for pediatric classification tasks.

Graph neural network architecture exploration experiments elucidated the impact of several architectural design choices on model performance and learning ability. Depth experiments revealed that increasing the number of GCN layers did not improve the simple GCN model. A moderate 3% accuracy improvement was observed on the pediatric stratification of the combined test set on increasing the depth of the residual GCN from 2 to 3 layers, but the further addition of a fourth GCN layer led to a performance drop on the pediatric stratification. These results indicate that for connectomic applications, a shallow depth of 2 to 3 GCN layers is sufficient for strong performance.

Architecture experiments also highlighted the importance of skip connections for the residual GCN model's ability to learn pediatric connectomic patterns. An ablation study removing skip connections resulted in a 14-point performance drop in pediatric classification accuracy. Because the only major distinction between the ablated residual network and the simple GCN model was the use of aggregation versus pooling in order to create an embedded graph representation, these results may suggest that connectomic models

which employ node aggregation approaches should introduce skip connections for optimal performance.

In addition to accuracy and AUC score metrics, models were assessed on adversarial robustness. The simple GCN model demonstrated the strongest adversarial robustness to FGSM attacks for larger perturbations. The residual GCN, on the other hand, demonstrated a notable lack of adversarial robustness, falling to below 40% classification accuracy for adversarial attacks with epsilon greater than or equal to 0.0005. The residual GCN model's lack of robustness may in part be a result of its higher complexity. Previous work has shown that higher complexity models exhibit greater sensitivity to adversarial attacks [55]. Further, training and validation loss curves illustrated that the residual GCN overfitted the training set. It is likely that the residual GCN model learned patterns derived from noise in the training set, making it more susceptible to adversarial attacks. Therefore, although the residual GCN model's higher complexity likely allowed it to learn more generalizable connectomic patterns than the other tested models, this complexity also resulted in a weakness to adversarial attacks. This potential tradeoff between raw test accuracy and adversarial robustness should be considered during model development, especially in the medical domain, where robustness is a key consideration for deploying trustworthy models.

While structural connectome data has recently been applied in graph classification tasks (e.g., sex classification) and graph prediction tasks (e.g., functional connectome prediction), graph regression tasks such as cognitive score prediction remain largely unexplored. These regression tasks hold potential for linking specific brain structural patterns to cognitive functions, thereby enhancing our understanding of cognitive processes and dysfunctions. Although both datasets selected for this study included measures of working memory and sustained attention for most participants, models developed here were not successful in predicting cognitive scores from structural connectome data. Future research with larger datasets or alternative learning approaches may provide better outcomes in this area.

5. Conclusions

This study demonstrated the effectiveness of GNNs in classifying sex based on structural connectome data, particularly in pediatric populations, a domain previously unexplored. Training and evaluating GNNs alongside standard machine learning models confirmed that GNNs outperformed other methods in both adult and pediatric datasets. While adult-trained models exhibit limited applicability to pediatric participants, enriching a pediatric dataset with adult data enables the development of a robust pediatric GNN model with performance comparable to adult models. A complex residual GCN architecture enables the best model performance in terms of classification accuracy and AUC score, while a simpler GCN architecture demonstrates greater adversarial robustness at the expense of a moderate performance drop. These findings underscored the potential of GNNs in advancing our understanding of sex-specific neurological development and disorders, highlighting important tradeoffs in the GNN architectural landscape and underscoring the importance of data enrichment in overcoming challenges associated with small pediatric datasets. Future research should continue to explore the application of GNNs to diverse clinical populations and investigate additional strategies for enhancing model generalizability across demographic groups.

Author Contributions: Conceptualization, A.S., R.R. and W.E.R.; methodology, A.S. and R.R.; software, A.S., R.R. and J.O.G.; validation, A.S., R.R. and W.E.R.; formal analysis, A.S., R.R. and W.E.R.; investigation, A.S., R.R., J.O.G. and W.E.R.; resources, M.M.H., N.D.S., K.R.K. and W.E.R.; data curation, A.S., R.R. and W.E.R.; writing—original draft preparation, A.S. and R.R.; writing—review

and editing, A.S., R.R., J.O.G., M.M.H., N.D.S., K.R.K. and W.E.R.; visualization, A.S. and W.E.R.; supervision, R.R. and W.E.R.; project administration, W.E.R.; funding acquisition, M.M.H., N.D.S., K.R.K. and W.E.R. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of St. Jude Children's Research Hospital (Protocol BRIGHT Revision 4.2, approved: 3 September 2017).

Informed Consent Statement: Patient consent was waived due to secondary use of research data exempted under Category 4 of the federal regulations because the information is recorded by the investigator in such a manner that the resulting dataset contains no information that can identify subjects, directly or through identifiers linked to the subjects.

Data Availability Statement: Pediatric dHCP data is publicly available. All code used in this study is publicly available: https://github.com/SrinivasanAnand/GNN_structural_connectivity.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Griffa, A.; Baumann, P.S.; Thiran, J.P.; Hagmann, P. Structural connectomics in brain diseases. *Neuroimage* **2013**, *80*, 515–526. https://doi.org/10.1016/j.neuroimage.2013.04.056.

2. Dennis, E.L.; Jahanshad, N.; McMahon, K.L.; de Zubicaray, G.I.; Martin, N.G.; Hickie, I.B.; Toga, A.W.; Wright, M.J.; Thompson, P.M. Development of brain structural connectivity between ages 12 and 30: A 4-Tesla diffusion imaging study in 439 adolescents and adults. *Neuroimage* **2013**, *64*, 671–684. https://doi.org/10.1016/j.neuroimage.2012.09.004.

3. Somerville, L.H.; Bookheimer, S.Y.; Buckner, R.L.; Burgess, G.C.; Curtiss, S.W.; Dapretto, M.; Elam, J.S.; Gaffrey, M.S.; Harms, M.P.; Hodge, C.; et al. The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5-21 year olds. *Neuroimage* **2018**, *183*, 456–468. https://doi.org/10.1016/j.neuroimage.2018.08.050.

4. Bernhardt, B.C.; Chen, Z.; He, Y.; Evans, A.C.; Bernasconi, N. Graph-theoretical analysis reveals disrupted small-world organization of cortical thickness correlation networks in temporal lobe epilepsy. *Cereb. Cortex* **2011**, *21*, 2147–2157. https://doi.org/10.1093/cercor/bhq291.

5. Gerster, M.; Taher, H.; Skoch, A.; Hlinka, J.; Guye, M.; Bartolomei, F.; Jirsa, V.; Zakharova, A.; Olmi, S. Patient-Specific Network Connectivity Combined with a Next Generation Neural Mass Model to Test Clinical Hypothesis of Seizure Propagation. *Front. Syst. Neurosci.* **2021**, *15*, 675272. https://doi.org/10.3389/fnsys.2021.675272.

6. Jirsa, V.K.; Proix, T.; Perdikis, D.; Woodman, M.M.; Wang, H.; Gonzalez-Martinez, J.; Bernard, C.; Benar, C.; Guye, M.; Chauvel, P.; et al. The Virtual Epileptic Patient: Individualized whole-brain models of epilepsy spread. *Neuroimage* **2017**, *145*, 377–388. https://doi.org/10.1016/j.neuroimage.2016.04.049.

7. Zhang, Y.; Lin, L.; Lin, C.P.; Zhou, Y.; Chou, K.H.; Lo, C.Y.; Su, T.P.; Jiang, T. Abnormal topological organization of structural brain networks in schizophrenia. *Schizophr. Res.* **2012**, *141*, 109–118. https://doi.org/10.1016/j.schres.2012.08.021.

8. Skudlarski, P.; Jagannathan, K.; Anderson, K.; Stevens, M.C.; Calhoun, V.D.; Skudlarska, B.A.; Pearlson, G. Brain connectivity is not only lower but different in schizophrenia: A combined anatomical and functional approach. *Biol. Psychiatry* **2010**, *68*, 61–69. https://doi.org/10.1016/j.biopsych.2010.03.035.

9. Yao, Z.; Zhang, Y.; Lin, L.; Zhou, Y.; Xu, C.; Jiang, T.; Alzheimer's Disease Neuroimaging Initiative. Abnormal cortical networks in mild cognitive impairment and Alzheimer's disease. *PLoS Comput. Biol.* **2010**, *6*, e1001006. https://doi.org/10.1371/journal.pcbi.1001006.

10. Wang, J.; Khosrowabadi, R.; Ng, K.K.; Hong, Z.; Chong, J.S.X.; Wang, Y.; Chen, C.Y.; Hilal, S.; Venketasubramanian, N.; Wong, T.Y.; et al. Alterations in Brain Network Topology and Structural-Functional Connectome Coupling Relate to Cognitive Impairment. *Front. Aging Neurosci.* **2018**, *10*, 404. https://doi.org/10.3389/fnagi.2018.00404.

11. Lo, C.Y.; Wang, P.N.; Chou, K.H.; Wang, J.; He, Y.; Lin, C.P. Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in Alzheimer's disease. *J. Neurosci.* **2010**, *30*, 16876–16885. https://doi.org/10.1523/JNEUROSCI.4136-10.2010.

12. Tolan, E.; Isik, Z. Graph Theory Based Classification of Brain Connectivity Network for Autism Spectrum Disorder. In *Bioinformatics and Biomedical Engineering: 6th International Work-Conference, IWBBIO 2018, Granada, Spain, 25–27 April 2018, Proceedings, Part I 6*; Spinger: Cham, Switzerland, 2018; pp. 520–530.

13. Li, H.; Xue, Z.; Ellmore, T.M.; Frye, R.E.; Wong, S.T. Network-based analysis reveals stronger local diffusion-based connectivity and different correlations with oral language skills in brains of children with high functioning autism spectrum disorders. *Hum. Brain Mapp.* **2014**, *35*, 396–413. https://doi.org/10.1002/hbm.22185.

14. Tymofiyeva, O.; Yuan, J.P.; Huang, C.Y.; Connolly, C.G.; Henje Blom, E.; Xu, D.; Yang, T.T. Application of machine learning to structural connectome to predict symptom reduction in depressed adolescents with cognitive behavioral therapy (CBT). *Neuroimage Clin.* **2019**, *23*, 101914. https://doi.org/10.1016/j.nicl.2019.101914.

15. Kamiya, K.; Amemiya, S.; Suzuki, Y.; Kunii, N.; Kawai, K.; Mori, H.; Kunimatsu, A.; Saito, N.; Aoki, S.; Ohtomo, K. Machine Learning of DTI Structural Brain Connectomes for Lateralization of Temporal Lobe Epilepsy. *Magn. Reson. Med. Sci.* **2016**, *15*, 121–129. https://doi.org/10.2463/mrms.2015-0027.

16. Chen, V.C.; Lin, T.Y.; Yeh, D.C.; Chai, J.W.; Weng, J.C. Functional and Structural Connectome Features for Machine Learning Chemo-Brain Prediction in Women Treated for Breast Cancer with Chemotherapy. *Brain Sci.* **2020**, *10*, 851. https://doi.org/10.3390/brainsci10110851.

17. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.

18. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2009**, *20*, 61–80. https://doi.org/10.1109/TNN.2008.2005605.

19. Bessadok, A.; Mahjoub, M.A.; Rekik, I. Graph Neural Networks in Network Neuroscience. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 5833–5848. https://doi.org/10.1109/TPAMI.2022.3209686.

20. Kazi, A.; Mora, J.; Fischl, B.; Dalca, A.V.; Aganj, I. Multi-Head Graph Convolutional Network for Structural Connectome Classification. In *Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology: 5th MICCAI Workshop, GRAIL 2023 and 1st MICCAI Challenge, OCELOT 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, 23 September, and 4 Octobe 2023, Proceedings*; Springer: Berlin/Heidelberg, Germany, 2024; Volume 14373, pp. 27–36. https://doi.org/10.1007/978-3-031-55088-1_3.

21. Weis, S.; Patil, K.R.; Hoffstaedter, F.; Nostro, A.; Yeo, B.T.T.; Eickhoff, S.B. Sex Classification by Resting State Brain Connectivity. *Cereb. Cortex* **2020**, *30*, 824–835. https://doi.org/10.1093/cercor/bhz129.

22. Gur, R.E.; Gur, R.C. Sex differences in brain and behavior in adolescence: Findings from the Philadelphia Neurodevelopmental Cohort. *Neurosci. Biobehav. Rev.* **2016**, *70*, 159–170. https://doi.org/10.1016/j.neubiorev.2016.07.035.

23. Beacher, F.D.; Minati, L.; Baron-Cohen, S.; Lombardo, M.V.; Lai, M.C.; Gray, M.A.; Harrison, N.A.; Critchley, H.D. Autism attenuates sex differences in brain structure: A combined voxel-based morphometry and diffusion tensor imaging study. *AJNR Am. J. Neuroradiol.* **2012**, *33*, 83–89. https://doi.org/10.3174/ajnr.A2880.

24. Williamson, J.; Yabluchanskiy, A.; Mukli, P.; Wu, D.H.; Sonntag, W.; Ciro, C.; Yang, Y. Sex differences in brain functional connectivity of hippocampus in mild cognitive impairment. *Front. Aging Neurosci.* **2022**, *14*, 959394. https://doi.org/10.3389/fnagi.2022.959394.

25. Zhang, J.; Gao, J.; Shi, H.; Huang, B.; Wang, X.; Situ, W.; Cai, W.; Yi, J.; Zhu, X.; Yao, S. Sex differences of uncinate fasciculus structural connectivity in individuals with conduct disorder. *Biomed. Res. Int.* **2014**, *2014*, 673165. https://doi.org/10.1155/2014/673165.

26. Tyan, Y.S.; Liao, J.R.; Shen, C.Y.; Lin, Y.C.; Weng, J.C. Gender differences in the structural connectome of the teenage brain revealed by generalized q-sampling MRI. *Neuroimage Clin.* **2017**, *15*, 376–382. https://doi.org/10.1016/j.nicl.2017.05.014.

27. Said, A.; Bayrak, R.G.; Derr, T.; Shabbir, M.; Moyer, D.; Chang, C.; Koutsoukos, X. NeuroGraph: Benchmarks for graph machine learning in brain connectomics. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.

28. Supekar, K.; Uddin, L.Q.; Prater, K.; Amin, H.; Greicius, M.D.; Menon, V. Development of functional and structural connectivity within the default mode network in young children. *Neuroimage* **2010**, *52*, 290–301. https://doi.org/10.1016/j.neuroimage.2010.04.009.

29. Mitchell, W.J.; Tepfer, L.J.; Henninger, N.M.; Perlman, S.B.; Murty, V.P.; Helion, C. Developmental Differences in Affective Representation Between Prefrontal and Subcortical Structures. *Soc. Cogn. Affect. Neurosci.* **2021**, *17*, 311–322. https://doi.org/10.1093/scan/nsab093.

30. Smith, S.M.; Jenkinson, M.; Woolrich, M.W.; Beckmann, C.F.; Behrens, T.E.; Johansen-Berg, H.; Bannister, P.R.; De Luca, M.; Drobnjak, I.; Flitney, D.E.; et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **2004**, *23* (Suppl. S1), S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051.

31. Tournier, J.D.; Smith, R.; Raffelt, D.; Tabbara, R.; Dhollander, T.; Pietsch, M.; Christiaens, D.; Jeurissen, B.; Yeh, C.H.; Connelly, A. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage.* 2019, 202, 116137. doi:10.1016/j.neuroimage.2019.116137.

32. Smith, R.E.; Tournier, J.D.; Calamante, F.; Connelly, A. SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *Neuroimage.* 2015, 119, 338-351. doi:10.1016/j.neuroimage.2015.06.092.

33. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM. A multi-modal parcellation of human cerebral cortex. *Nature.* **2016**, *536*, 171–178. https://doi.org/10.1038/nature18933.

34. Kipf, T.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907. https://doi.org/10.48550/arXiv.1609.02907.

35. Jernigan, T.L.; Brown, T.T.; Hagler, D.J., Jr.; Akshoomoff, N.; Bartsch, H.; Newman, E.; Thompson, W.K.; Bloss, C.S.; Murray, S.S.; Schork, N.; et al. The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository. *Neuroimage* **2016**, *124*, 1149–1154. https://doi.org/10.1016/j.neuroimage.2015.04.057.

36. Casey, B.J.; Cannonier, T.; Conley, M.I.; Cohen, A.O.; Barch, D.M.; Heitzeg, M.M.; Soules, M.E.; Teslovich, T.; Dellarco, D.V.; Garavan, H.; et al. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **2018**, *32*, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001.

37. Bullmore, E.; Sporns, O. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **2009**, *10*, 186–198. https://doi.org/10.1038/nrn2575.

38. Hagmann, P.; Cammoun, L.; Gigandet, X.; Meuli, R.; Honey, C.J.; Wedeen, V.J.; Sporns, O. Mapping the structural core of human cerebral cortex. *PLoS Biol.* **2008**, *6*, e159. https://doi.org/10.1371/journal.pbio.0060159.

39. van den Heuvel, M.P.; Sporns, O. Network hubs in the human brain. *Trends Cogn. Sci.* **2013**, *17*, 683–696. https://doi.org/10.1016/j.tics.2013.09.012.

40. Tomasi, D.; Volkow, N.D. Functional connectivity hubs in the human brain. *Neuroimage* **2011**, *57*, 908–917. https://doi.org/10.1016/j.neuroimage.2011.05.024.

41. Collin, G.; Sporns, O.; Mandl, R.C.; van den Heuvel, M.P. Structural and functional aspects relating to cost and benefit of rich club organization in the human cerebral cortex. *Cereb. Cortex* **2014**, *24*, 2258–2267. https://doi.org/10.1093/cercor/bht064.

42. Hofer, S.; Frahm, J. Topography of the human corpus callosum revisited—Comprehensive fiber tractography using diffusion tensor magnetic resonance imaging. *Neuroimage* **2006**, *32*, 989–994. https://doi.org/10.1016/j.neuroimage.2006.05.044.

43. Wahl, M.; Lauterbach-Soon, B.; Hattingen, E.; Jung, P.; Singer, O.; Volz, S.; Klein, J.C.; Steinmetz, H.; Ziemann, U. Human motor corpus callosum: Topography, somatotopy, and link between microstructure and function. *J. Neurosci.* **2007**, *27*, 12132–12138. https://doi.org/10.1523/JNEUROSCI.2320-07.2007.

44. Lebel, C.; Caverhill-Godkewitsch, S.; Beaulieu, C. Age-related regional variations of the corpus callosum identified by diffusion tensor tractography. *Neuroimage* **2010**, *52*, 20–31. https://doi.org/10.1016/j.neuroimage.2010.03.072.

45. Jancke, L.; Staiger, J.F.; Schlaug, G.; Huang, Y.; Steinmetz, H. The relationship between corpus callosum size and forebrain volume. *Cereb. Cortex* **1997**, *7*, 48–56. https://doi.org/10.1093/cercor/7.1.48.

46. Ingalhalikar, M.; Smith, A.; Parker, D.; Satterthwaite, T.D.; Elliott, M.A.; Ruparel, K.; Hakonarson, H.; Gur, R.E.; Gur, R.C.; Verma, R. Sex differences in the structural connectome of the human brain. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 823–828. https://doi.org/10.1073/pnas.1316909110.

47. Gilmore, J.H.; Knickmeyer, R.C.; Gao, W. Imaging structural and functional brain development in early childhood. *Nat. Rev. Neurosci.* **2018**, *19*, 123–137. https://doi.org/10.1038/nrn.2018.1.

48. Cao, M.; Huang, H.; He, Y. Developmental Connectomics from Infancy through Early Childhood. *Trends Neurosci.* **2017**, *40*, 494–506. https://doi.org/10.1016/j.tins.2017.06.003.

49. Sydnor, V.J.; Larsen, B.; Bassett, D.S.; Alexander-Bloch, A.; Fair, D.A.; Liston, C.; Mackey, A.P.; Milham, M.P.; Pines, A.; Roalf, D.R.; et al. Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology. *Neuron* **2021**, *109*, 2820–2846. https://doi.org/10.1016/j.neuron.2021.06.016.

50. Margulies, D.S.; Ghosh, S.S.; Goulas, A.; Falkiewicz, M.; Huntenburg, J.M.; Langs, G.; Bezgin, G.; Eickhoff, S.B.; Castellanos, F.X.; Petrides, M.; et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 12574–12579. https://doi.org/10.1073/pnas.1608282113.

51. Hagmann, P.; Sporns, O.; Madan, N.; Cammoun, L.; Pienaar, R.; Wedeen, V.J.; Meuli, R.; Thiran, J.P.; Grant, P.E. White matter maturation reshapes structural connectivity in the late developing human brain. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 19067–19072. https://doi.org/10.1073/pnas.1009073107.

52. Dennis, E.L.; Thompson, P.M. Typical and atypical brain development: A review of neuroimaging studies. *Dialogues Clin. Neurosci.* **2013**, *15*, 359–384. https://doi.org/10.31887/DCNS.2013.15.3/edennis.

53. Lebel, C.; Beaulieu, C. Longitudinal development of human brain wiring continues from childhood into adulthood. *J. Neurosci.* **2011**, *31*, 10937–10947. https://doi.org/10.1523/JNEUROSCI.5302-10.2011.

54. Westlye, L.T.; Walhovd, K.B.; Dale, A.M.; Bjornerud, A.; Due-Tonnessen, P.; Engvig, A.; Grydeland, H.; Tamnes, C.K.; Ostby, Y.; Fjell, A.M. Life-span changes of the human brain white matter: Diffusion tensor imaging (DTI) and volumetry. *Cereb. Cortex* **2010**, *20*, 2055–2068. https://doi.org/10.1093/cercor/bhp280.

55. Rodriguez, D.; Nayak, T.; Chen, Y.; Krishnan, R.; Huang, Y. On the role of deep learning model complexity in adversarial robustness for medical images. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 160. https://doi.org/10.1186/s12911-022-01891-w.