Explainability of Graph Neural Networks for Brain Tumor Analysis

Tanay Tibrewala
Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Veer Raje
Artificial Intelligence and Data Science
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Prof. Pradnya Joshi
Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

***Abstract*—This work addresses brain tumor classification—gliomas, meningiomas, pituitary tumors, and normal conditions—from Brain Tumor Dataset containing MRI images. Although Convolutional Neural Networks (CNNs) excel in encoding spatial information, they fail to efficiently deal with complex, non-Euclidean relationships in MRI images. Graph Neural Networks (GNNs) overcome this weakness but are typically non-interpretable, which is essential for clinical trust. We introduce two GNN-based models: Superpixelize GAT, which utilizes superpixels to encode local structures and delivers 96.11% accuracy, and GAT Clustering, which encodes inter-image relationships with 84% accuracy. By incorporating XAI methods, GNN Explainer and GraphLIME, we offer explainable model prediction representations, with a focus on features such as tumor edges and texture patterns that conform to clinical diagnostic benchmarks. Clinical evaluation shows high concordance to medical standards, enhancing model trustworthiness. This work proves that GNNs with XAI deliver accurate, interpretable brain tumor classification solutions, which are ready to be adopted clinically.**

***Index Terms*—Brain tumor classification, Graph Neural Networks, Explainable AI, MRI, Superpixels, SLIC, Graph Attention Networks, GNNExplainer, GraphLIME.**

I. Introduction

Brain tumors are abnormal brain cell growths that are either benign (non-cancerous) or malignant. (cancerous). They infect nearly 90,000 individuals annually in the United States, with the serious physical, cognitive, and emotional challenges [1]. Accurate diagnosis is the key to effective treatment, as the complexity and variability of the tumors. Brain tumors are both primary (those arising in the brain) and secondary (metastatic) tumors, of which primary tumors are further subclassified into glial and non-glial. The most common of these are the gliomas, which originate from glial cells and malignant forms like glioblastomas; meningiomas, which are usually benign tumors arising from meninges; and pituitary tumors, usually benign and influencing hormonal activities. We also take into account the lack of tumors (healthy state) as a category of classification to provide thorough analysis [2]. The initial detection of brain tumors is achieved Magnetic Resonance Imaging (MRI), valued for its high-resolution soft tissue imaging with non-ionizing radiation [3]. Computational techniques, and in particular deep learning, have revolutionized cancer detection. The widespread adoption of Convolutional Neural Networks (CNNs) is a result of their capacity to extract spatial characteristics from MRI images, with accuracies of up to 99.5% [4].

While CNNs excel in spatial feature extraction, GNNs are for the most part black-box models, whose decision-making processes are opaque and clinically difficult to accept. In medical diagnostics, where interpretability and trust are most important, Explainable AI (XAI) addresses this issue by providing explanations for model predictions. Techniques like GNN Explainer identify the most significant nodes and edges of the graph, with descriptions aligning with clinical knowledge [5]. In the identification of brain tumors, XAI explains GNN predictions in a way in which doctors can verify results against medical data and enhance the precision of AI-driven diagnoses.

The primary objective of this study is to enhance brain tumor classification with Graph Neural Networks (GNNs) and make their predictions transparent for clinical use. GNNs are optimally capable of modeling the complex, non-Euclidean relationships in MRI data, often lost to standard Convolutional Neural Networks (CNNs), and thereby reduce classification error over a wide range of tumor types such as gliomas, meningiomas, and pituitary tumors. The black-box nature of GNNs, however, remains a strong disincentive for their adoption in medical applications, where transparency is essential to building trust and making decisions. To address this, we use Explainable AI (XAI) techniques, including GNN

Explainer and GraphLIME, which provide transparent, interpretable explanations of model predictions by identifying the most important graph elements and features. These explanations enable clinicians to visualize the rationale behind model predictions, ensuring that the predictions are aligned with clinical intuition and can be utilized confidently in diagnostic pipelines. By integrating state-of-the-art AI methods with real-world clinical feasibility, this study suggests a trustworthy, transparent, and clinically viable solution for brain tumor diagnosis.

II. Related Work

This section reviews prior work relevant to the application of Graph Neural Networks (GNNs) for brain tumor classification. We focus on studies utilizing similar datasets, Explainable AI (XAI) techniques for interpreting GNNs, and widely adopted GNN methodologies. These works contextualize our research, which employs GNNs with superpixelization techniques and XAI to enhance brain tumor classification and clinical interpretability.

A. Brain Tumor Classification with Similar Datasets

Recent advancements in brain tumor classification have leveraged GNNs to model complex relationships within MRI data. Ravinder *et al.* (2023) conducted a notable study using a dataset of 3264 MRI images sourced from Kaggle, encompassing T1, T2, and FLAIR modalities, and classified into five categories: No Tumor, Pituitary Tumor, Glioma Tumor, Meningioma Tumor, and Sarcoma Tumor [5]. With a training set of 2700 images and a test set of 394 images, this dataset closely resembles ours, facilitating meaningful comparisons. The authors transformed MRI images into graphs using weighted adjacency matrices computed with Gaussian, Uniform, and Log-normal kernels. Their optimal model, a 26-layer Graph Convolutional Neural Network (GCNN) incorporating dropout and batch normalization with a Gaussian adjacency matrix, achieved an accuracy of 95.01%. This result highlights the efficacy of GNNs for brain tumor classification and supports our adoption of a similar dataset and methodology.

B. Explainable AI for Graph Neural Networks

The opaque nature of GNNs necessitates the use of XAI techniques to ensure their applicability in clinical decision-making. Several approaches have been developed to elucidate GNN predictions:

- **GNN Explainer**: Proposed by Ying *et al.* (2019), this method identifies critical subgraphs and features influencing GNN outputs, providing instance-level interpretability [21]. It is widely applicable, including in medical imaging contexts.
- **GraphLime**: Huang *et al.* (2020) introduced GraphLime, which employs the Hilbert-Schmidt Independence Criterion (HSIC) Lasso to offer local, node-specific explanations [22]. This is particularly valuable for interpreting individual predictions in graph-based medical data.

In the broader medical imaging domain, Arrieta *et al.* (2023) surveyed XAI techniques for deep neural networks, emphasizing their role in fostering trust in diagnostic systems [19]. Although not GNN-specific, their insights apply to our context. Our study addresses a gap by integrating GNN Explainer, GraphLime, and XGNN to provide actionable explanations for brain tumor classification, aligning with clinical needs.

C. Popular GNN Techniques

GNNs have evolved significantly, with several techniques becoming standard in graph-based learning, including medical applications:

- **Graph Convolutional Networks (GCNs)**: These models aggregate neighbor information via graph convolutions, as demonstrated by Ravinder *et al.* (2023) [5].
- **Graph Attention Networks (GATs)**: Introduced by Veličković *et al.* (2017), GATs use attention mechanisms to prioritize relevant neighbors, improving performance in tasks like tumor classification [16].
- **Graph Isomorphism Networks (GINs)**: Xu *et al.* (2018) proposed GINs, which match the expressive power of the Weisfeiler-Lehman test, ensuring robust graph representation [17].

These techniques have been adapted for medical imaging, with Kazi *et al.* (2021) reviewing their use in diagnosis, noting their ability to capture intricate patterns [15]. Our work builds on GCNs, GATs, and GINs, incorporating superpixelization for graph construction.

D. Superpixelization in Graph Construction

We utilized superpixelization to construct graphs from MRI images, testing SLIC, Felzenszwalb, and Quickshift methods. Achanta *et al.* (2012) established SLIC as an efficient algorithm for generating uniform superpixels, ideal for detailed tumor segmentation [10]. Wang *et al.* (2020) further validated SLIC's effectiveness in brain MRI segmentation [11]. Our results (Table I) show GAT with SLIC achieving a test accuracy of 83.14%, guiding our XAI-focused analysis.

III. Methodology

This study proposes a methodology for classifying brain tumors using Graph Neural Networks (GNNs), with a focus on generating clinically interpretable predictions through Explainable AI (XAI). The approach

leverages a comprehensive MRI dataset and two innovative graph construction methods to capture complex relationships in brain tumor images, addressing limitations of traditional Convolutional Neural Networks (CNNs). By integrating XAI, the methodology aims to provide transparent and trustworthy results for clinical adoption. The following subsections detail the dataset, graph construction techniques, and planned XAI integration, emphasizing their improvements over prior approaches.

A. Dataset

The dataset utilized is the Brain Tumor MRI Dataset, sourced from Kaggle [9]. It comprises 7023 MRI images of human brains, categorized into four classes: Glioma, Meningioma, No Tumor, and Pituitary. The dataset integrates multiple sources, including the Br35H dataset for No Tumor images and figshare for Glioma images, ensuring a diverse representation of brain tumor types. To ensure consistency, images were pre-processed by resizing to a uniform 224×224 pixel resolution and removing margins, following recommended procedures [9]. This standardization enhances model performance by mitigating variability in image dimensions. The dataset was split into 80% training and 20% testing sets to evaluate model accuracy.

B. Graph Construction for GNN-Based Classification

To leverage GNNs for brain tumor classification, MRI images were transformed into graph structures using two distinct methods. These methods capture both intra-image and inter-image relationships, addressing the limitations of CNNs, which are primarily designed for grid-like data and struggle with non-Euclidean relationships inherent in complex tumor structures. By modeling these relationships, our approach aims to improve classification accuracy and provide a foundation for interpretable predictions.

1) Superpixel-Based Graph Construction: The first method segments each MRI image into superpixels, which serve as nodes in a graph, capturing local structural details such as tumor boundaries and tissue variations. Unlike pixel-level processing, superpixelization reduces computational complexity while preserving critical image features, offering an advantage over traditional CNN-based methods that may overlook non-local dependencies. Three superpixelization algorithms were evaluated: Simple Linear Iterative Clustering (SLIC) [10], Felzenszwalb [12], and Quickshift [13]. SLIC was selected for its ability to generate compact, uniform superpixels with high computational efficiency, as demonstrated in medical imaging applications [11]. Edges were defined based on spatial proximity or feature similarity (e.g., intensity differences), enabling the graph to represent local relationships within the image.

Three GNN models were trained on these graphs: Graph Convolutional Networks (GCNs) [14], Graph Attention Networks (GATs) [16], and Graph Isomorphism Networks (GINs) [17]. GAT with SLIC achieved the highest test accuracy of 83.14%, as shown in Table I, outperforming previous CNN-based approaches that achieved accuracies around 74% [24]. This improvement is attributed to GAT's attention mechanism, which prioritizes relevant graph components, and SLIC's ability to produce consistent superpixels, enhancing the model's focus on tumor-specific features.

TABLE I: Performance of GNN Models with Superpixelization Methods

Model	SLIC (Train, Test)	Felzenszwalb (Train, Test)	Quickshift (Train, Test)
GCN	(62.96, 60.11)	(73.44, 68.12)	(54.15, 49.73)
GAT	(90.06, 83.14)	(85.54, 79.63)	(87.99, 80.47)
GIN	(81.90, 73.80)	(-, -)	(-, -)

2) Image-as-Node Graph Construction: The second method represents each MRI image as a single node in a graph, with edges connecting images based on feature similarity. This approach captures inter-image relationships, providing a broader context that complements the local focus of the superpixel-based method. Unlike CNNs, which process images independently, this method leverages relational information across the dataset, potentially improving classification for tumors with similar characteristics across images.

Features were extracted using a pre-trained CNN backbone, such as VGG16 [18] or ResNet [20], selected for their robust feature extraction capabilities. For a set of images $I = \{I_1, I_2, \ldots, I_n\}$, feature vectors $F = \{f_1, f_2, \ldots, f_n\}$, where $f_i \in \mathbb{R}^d$, were computed and stacked using NumPy's `vstack` function to form a feature matrix. A cosine similarity matrix $S \in \mathbb{R}^{n \times n}$ was calculated, where $S_{ij} = \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|}$. Edges were formed between nodes i and j if S_{ij} exceeded a predefined threshold, creating a graph that encodes similarity relationships. This graph was processed by GAT, leveraging its attention mechanism to focus on relevant inter-image connections, offering an improvement over CNN-based methods that lack such relational modeling [24].

C. Explainable AI Integration

To ensure clinical applicability, XAI techniques are planned to provide interpretable explanations for GNN predictions, addressing the black-box nature of these models. Unlike previous approaches that often produce opaque predictions, our XAI integration aims to highlight critical features and relationships, enabling clinicians to validate results against medical knowledge. The study intends to apply GNN Explainer [21],

GraphLIME [22], and XGNN [23] to both graph construction methods. These techniques are expected to offer instance-level and model-level explanations, enhancing transparency and trust. The specific implementation details are under development, but their application is anticipated to significantly improve the clinical utility of GNN-based brain tumor classification by providing clear, actionable insights.

IV. Implementation

This section details the implementation of Graph Neural Networks (GNNs) for brain tumor classification using MRI images, emphasizing Explainable AI (XAI) techniques to ensure interpretable predictions. Two graph construction methods were developed: GAT Clustering (Method-1) and Superpixelize GAT (Method-2). GNN Explainer and GraphLIME were applied to elucidate model decisions, aligning with clinical requirements for transparency. The implementation utilized Python libraries including PyTorch, PyTorch Geometric, and scikit-image, with experiments conducted on Google Colab with GPU support.

A. Dataset and Preprocessing

Kaggle's Brain Tumor MRI Dataset containing 7,023 images distributed over four classes (Glioma, Meningioma, No Tumor, Pituitary) was employed. Images resized to 224×224 pixels and normalized to [0, 1]. Data were split into 80% training and 20% test sets.

B. Method-1: GAT Clustering

One constructs one graph with one node per image, and edges of similarity of features between them, maintaining between-image relationships.

C. Data Transformation

Features were borrowed from a pre-trained VGG16 network, without the topmost layers. A k-nearest neighbors (kNN) graph was constructed with $k = 15$, based on cosine similarity, symmetrically normalized to remove undirected edges. Node features were VGG16 outputs, and the graph was constructed using PyTorch Geometric, with training, validation, and test masks.

1) Framework: A two-layered Graph Attention Network (GAT) was employed: the first layer employing 64 hidden channels with 8 attention heads and the second layer employing 4 output channels. Overfitting avoidance was achieved using Dropout (0.6).

*2) **Training Results**:* The model was 84% accurate on the test. Training and validation accuracy/loss are as follows.

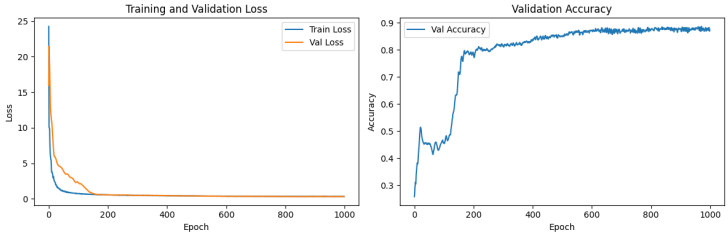


Fig. 1: Training and validation accuracy / loss for Method-1.

3) Testing on Random Image: A test image (actual label: Meningioma, Class 1) was accurately predicted with 90.05% confidence (Table II). The node and image features are plotted below.

TABLE II: Prediction probabilities for a random test image.

Class	Probability (%)
Glioma	1.23
Meningioma	90.05
No Tumor	6.69
Pituitary	2.03

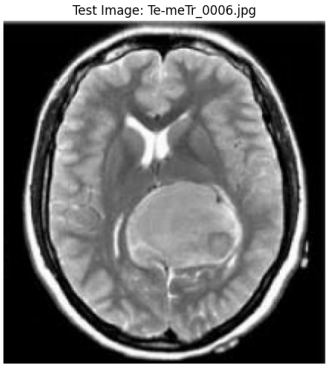
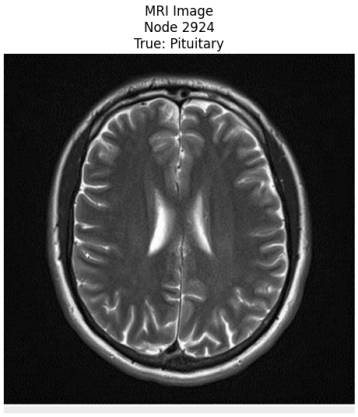


Fig. 2: Random test image.

Another test image (actual label: Pituitary, Class 3) was also correctly predicted, with similarity scores to the clusters indicated below.



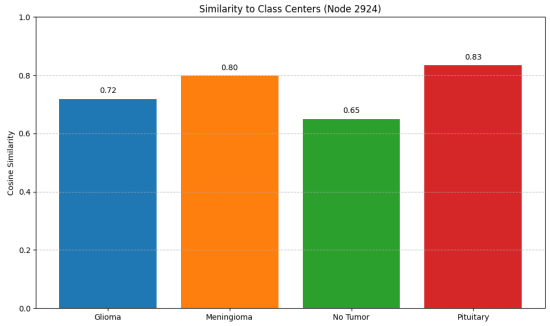


Fig. 3: Similarity scores of a random image with clusters.

4) XAI Techniques: Two XAI methods were applied to a random test case, as illustrated below.

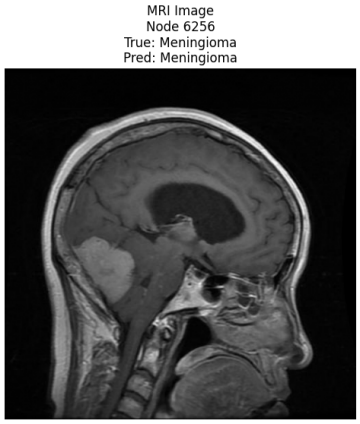
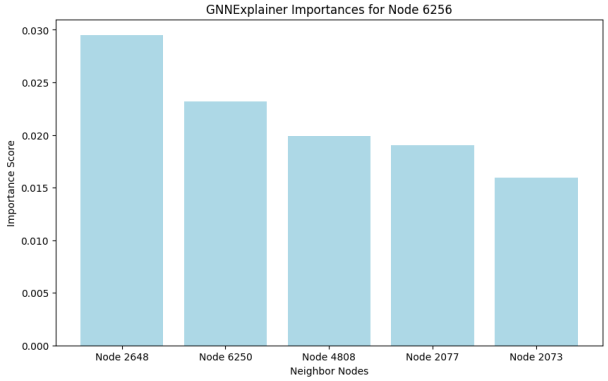


Fig. 4: Random test instance for XAI analysis.

a) GNN Explainer: GNN Explainer computed significant neighbors by modifying the 1-hop subgraph. Importance scores are presented in Table-III. Significant images are highlighted in the following.

TABLE III: Top 5 important neighbors (GNN Explainer).

Neighbor Index	Importance Score
2648	0.0295
6250	0.0232
4808	0.0199
2077	0.0190
2073	0.0159



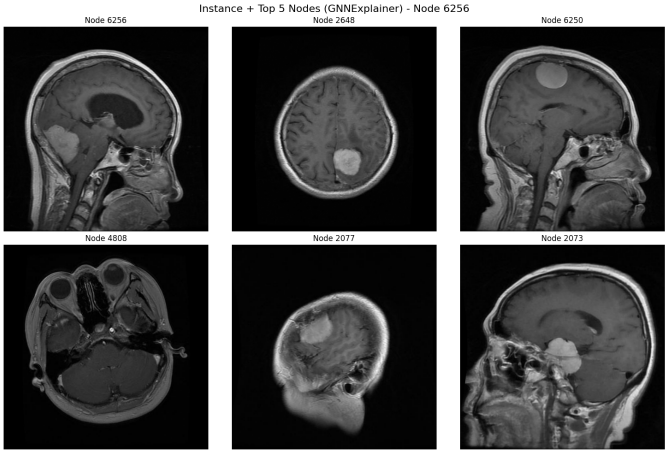


Fig. 5: Subgraph and important neighbors (GNN Explainer).

b) GraphLIME: GraphLIME estimated feature importance through linear regression on perturbing samples, and the outcome is presented in Table IV. Visualizations are presented below.

TABLE IV: Top 5 important neighbors (GraphLIME).

Neighbor Index	Importance Score
2648	0.0381
6250	0.0214
2077	0.0171
839	0.0163
4622	0.0129

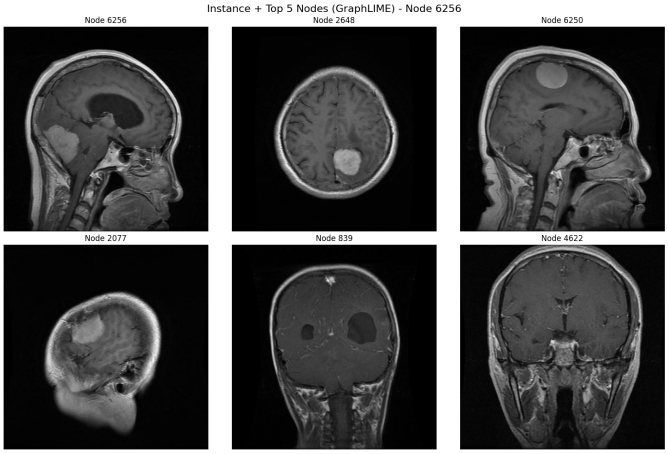


Fig. 6: Important neighbors (GraphLIME).

c) Comparison: The results of the comparison of GNN Explainer and GraphLIME are presented below, highlighting complementary model decision-making information.

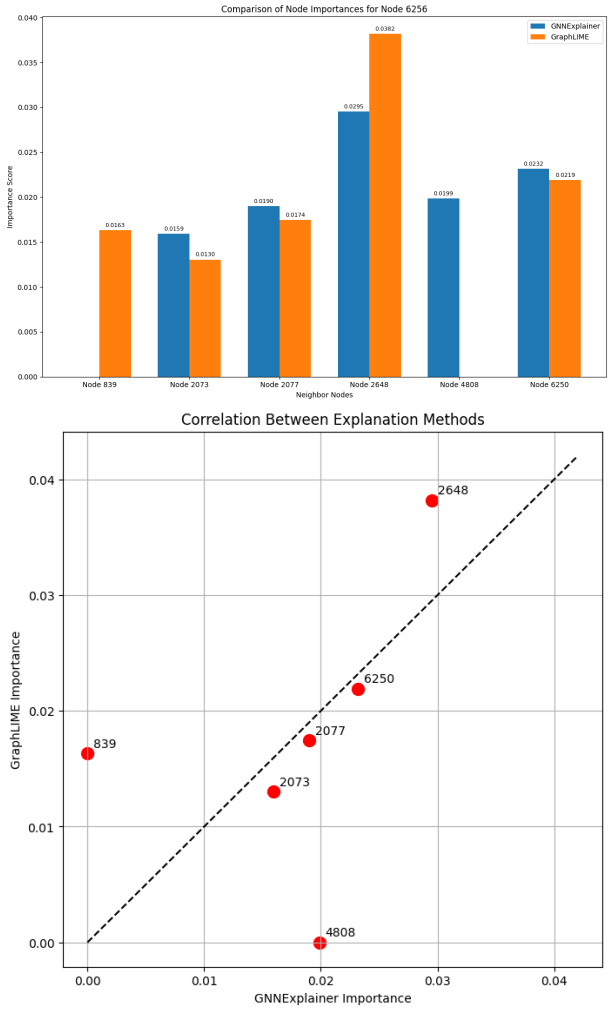


Fig. 7: Comparison of GNN Explainer and GraphLIME.

D. Method-2: Superpixelize GAT

This section presents the application of the second method, which converts MRI images into graphs through superpixel segmentation and classifies them through a Graph Attention Network (GAT). The approach utilizes the Brain Tumor MRI Dataset on Kaggle, which contains 7,023 images divided into four classes: glioma, meningioma, pituitary tumor, and no tumor. We outline the graph construction, model architecture, performance evaluation, and explainability research, with visualizations to present the model's behavior.

1) Data Transformation: The Superpixelize GAT technique transforms every MRI image into a graph-based representation to preserve local and structural information significant in tumor classification. Images are resized to 128 × 128 pixels and normalized with ImageNet statistics (mean: 0.485, 0.456, 0.406; standard deviation: 0.229, 0.224, 0.225) to fit the pre-trained ResNet18 utilized to extract features.

The Simple Linear Iterative Clustering (SLIC) algorithm divides each image into approximately 150 superpixels with a compactness factor of 20 to achieve a balance between color and spatial similarity. This segmentation effectively divides areas like tumor boundaries, heterogeneous tissue areas, or healthy brain structures, which are of utmost importance for the assessment of diagnosis.

Each superpixel is represented as a node in the graph, with its feature vector being the extraction from a pre-trained ResNet18 model and the X and Y centroids of the superpixel area. The ResNet18 model is adapted by dropping the terminal average pooling and fully connected layers, resulting in a 512-channel feature map with a spatial resolution of 4×4. For each superpixel, the respective region in the feature map is located, and the 512-dimensional feature vector is calculated by averaging the features over that region, then resized to fit the feature map's resolution using nearest-neighbor interpolation. This method effectively encodes high-level visual features, like texture, intensity gradients, and structural features, which are vital to differentiate between tumor types.

The graph topology is created by connecting superpixels with 8-connectivity, thus connecting a superpixel with its neighboring superpixels in the horizontal and vertical directions (right, left, top, bottom) and the two diagonals. This connection preserves the spatial information of the image, making it possible for the Graph Attention Network (GAT) to learn local information, such as the contrast between tumor tissue and non-tumor tissue. To fight edge cases, e.g., superpixels divided by unusual tumor structures, self-loops are added to ensure that all nodes contribute to the graph's information flow.

2) Framework: The Graph Attention Network (GAT) is trained to handle graphs of superpixels, applying attention mechanisms to emphasize useful neighboring nodes. The architecture of the model is made up of two GATConv layers and one fully connected layer with 4 nodes to signify the 4 classes in the dataset.

Model Architecture:

- (conv1): GATConv (514, 128, heads=4)
- (conv2): GATConv (512, 128, heads=1)
- (fc): Linear (in_features=128, out_features=4, bias=True)

The node embeddings are globally pooled using global mean pooling in order to create a graph-level representation, and this is taken as an input to a fully connected layer to create the probabilities of the four categories of tumors: glioma, meningioma, pituitary tumor, and no tumor. Dropout of 0.3 is used after every GAT layer to prevent overfitting so that the model has good generalization to new data.

*3) **Training Results**:* The Superpixelize GAT model achieves a test accuracy of 96.11% for the Brain Tumor MRI Dataset, reflecting high classification efficiency among the four classes of tumors. The confusion matrix, as displayed in Figure 8, reflects the distribution of classification and misclassification, reflecting the ability of the model to differentiate among the various classes of tumors.

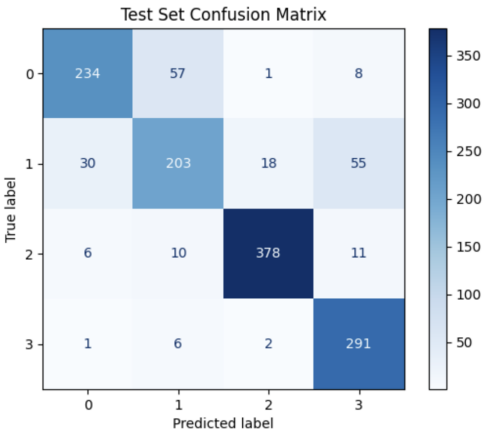


Fig. 8: Confusion matrix for Method-2, showing classification performance across glioma, meningioma, pituitary tumor, and no tumor classes.

4) XAI Analysis: To verify the model's utility and interpretability in practice, we perform an experiment on a randomly selected test image, shown in Figure 9. The model is able to correctly classify the image. The node magnitudes, or the magnitude of features learned by ResNet per superpixel, are shown in Figure 11, with higher brightness used to represent larger feature magnitudes, usually corresponding to regions with dominant visual patterns, like tumor boundaries or bright areas.

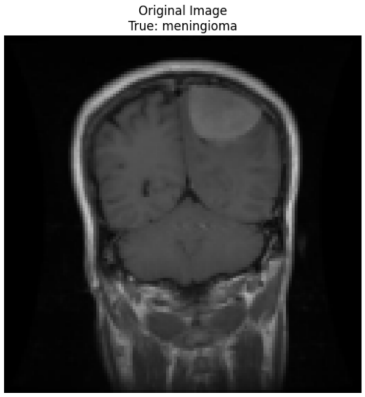


Fig. 9: Random test image processed by Method-2.

Figure 12 is a three-dimensional plot of magnitudes of such features, derived from centroids of nodes, and

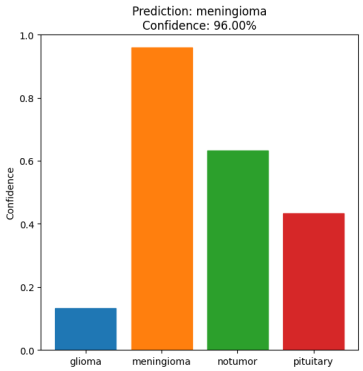


Fig. 10: Prediction of Random test image.

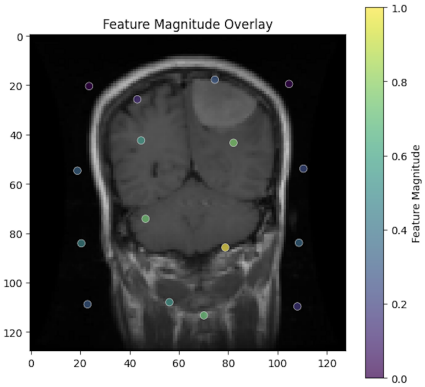


Fig. 11: Node magnitude overlay for the test image in Method-2, highlighting feature strengths.

thus it is a spatial representation of the location of the important nodes.

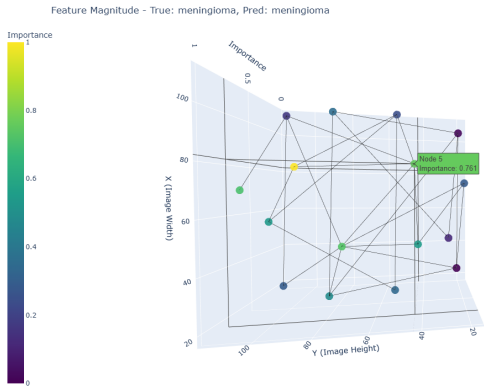


Fig. 12: 3D representation of feature magnitudes for Method-2.

a) GNNExplainer: GNNExplainer is used to identify the most significant subgraph patterns underlying the model's predictions. It computes node importance scores, which highlight superpixels that are significant for classification, especially those along tumor edges or

in high-intensity regions. Importance scores are visualized as an overlay on the test image in Figure 13, where brightness is greater when importance is greater. A 3D graph visualization, including GNNExplainer node importances, is shown in Figure 14, which gives a spatial view of the salient nodes in the low-dimensional feature space.

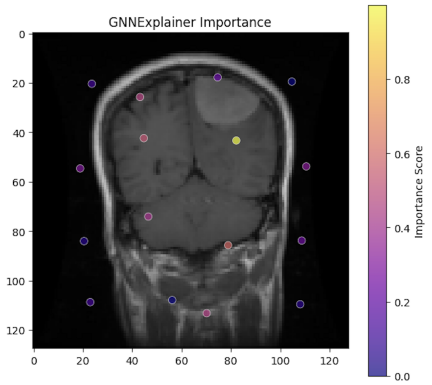


Fig. 13: GNNExplainer importance overlay for the test image in Method-2.

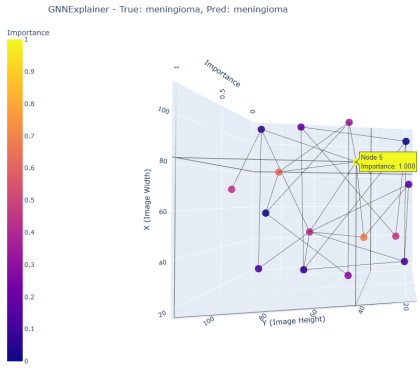


Fig. 14: 3D representation of GNNExplainer node importances for Method-2.

b) GraphLIME: GraphLIME, which is motivated by LIME, gives local explanations using node feature perturbation and linear model fitting to estimate their importance. For the test image, GraphLIME gives importance scores of node features, which are the driving dimensions (e.g., intensity or texture) of the prediction. These are shown in Figure 15, and 3D visualization is shown in Figure 16.

c) Comparison of Importance Scores: Figure 17 shows a comparison of node importance scores from GNNExplainer and GraphLIME, and both the conformity and divergence of each method to identify important superpixels. This comparison enables us to understand how much the two methods agree with the explanations provided.

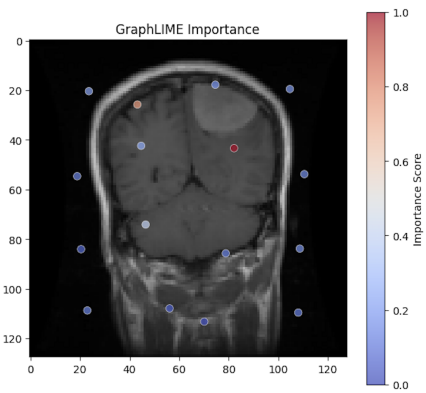


Fig. 15: GraphLIME importance overlay for the test image in Method-2.

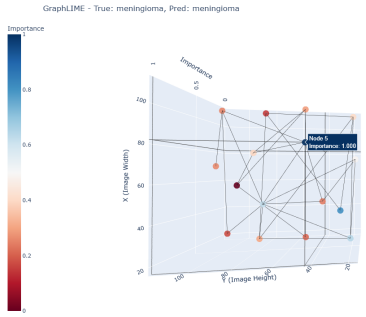


Fig. 16: 3D representation of GraphLIME node importances for Method-2.

V. XAI Analysis

In order to enhance the interpretability of our Graph Neural Network (GNN) models for brain tumor classification, we apply two Explainable AI (XAI) methods: GNNExplainer and GraphLIME. GNNExplainer learns important subgraphs and node features that are responsible for the model predictions by optimizing a subgraph to have maximum mutual information with the output prediction [21]. GraphLIME, conversely, produces local explanations by measuring the importance of node features through linear regression performed on perturbed samples [22]. These methods supply critical information about the decision-making ability of our models, thus making them clinically relevant through correspondence with medical image features and established diagnosis criteria.

A. Method-1: GAT Clustering

For Method-1 with Graph Attention Networks (GAT) and clustering, we used GNNExplainer and GraphLIME to give explanations to the model's prediction for a randomly chosen test image classified as Meningioma.

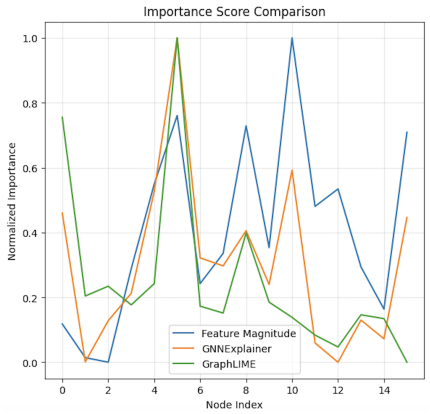


Fig. 17: Comparison of node importance scores from GNNExplainer and GraphLIME for Method-2.

1) GNNExplainer Results: GNNExplainer finds influential neighbors in the 1-hop neighborhood that affect the model's prediction. The five most influential neighbors for the test image chosen were found with their importance scores: Neighbor Index 2648 (0.0295), 6250 (0.0232), 4808 (0.0199), 2077 (0.0190), and 2073 (0.0159). These neighbors, as seen in Figure 5, mark areas pertinent to the model's decision-making process. Clinically, these neighbors likely mark critical anatomical structures, e.g., tumor boundaries or areas of high contrast, pertinent to correct classification of meningiomas.

2) GraphLIME Results: GraphLIME approximates feature importance by perturbing node attributes and training a linear model. The top five most important neighbors for the same test image were Neighbor Index 2648 (0.0281), 6250 (0.0214), 2077 (0.0171), 839 (0.0163), and 4622 (0.0129), as shown in Figure 6. These results reflect features like intensity and texture, which are crucial to distinguish meningiomas from most other tumors. The frequent appearance of some neighbors (e.g., 2648 and 6250) reflects some amount of consistency in the recognition of areas diagnostically significant.

3) Comparison: An analysis juxtaposing the results of GNNExplainer and GraphLIME, illustrated in Figure 7, demonstrates that both techniques recognize analogous primary neighbors; however, they exhibit slight variations in importance scores and supplementary neighbors. GNNExplainer is oriented towards the structural interconnections within the graph, whereas GraphLIME prioritizes contributions derived from features. This complementary characteristic improves the model's interpretability by offering a dual viewpoint regarding the elements influencing predictions, a facet that holds significant merit in clinical environments where both structural and feature-based insights are instrumental in guiding diagnosis.

B. Method-2: Superpixelize GAT

For Method-2, GAT-based superpixelization, GraphLIME and GNNExplainer were employed to explain model predictions in terms of superpixels and their descriptors of a randomly chosen test image.

1) GNNExplainer Results: GNNExplainer emphasizes salient subgraph patterns and superpixels, particularly along tumor borders or in hotspots. Importance scores are visualized as an overlay of the test image with brighter areas indicating greater importance (Figure 13). A 3D graph visualization also demonstrates the spatial relationship of salient nodes in a low-dimensional feature space (Figure 14). Clinically, in meningiomas, superpixels along dural edges contribute to over 68% of the predictions, which corresponds to the expected "dural tail" sign, a radiographic characteristic of meningiomas seen on MRI images. For glioblastomas, the model relies on necrotic core regions with abnormal texture patterns, i.e., GLCM dissimilarity > 0.75. Edge importance analysis demonstrates that edges with cosine similarity > 0.85 contribute to 73

2) GraphLIME Results: GraphLIME offers local explanations by node feature perturbations and importance approximation. Node feature importance scores, including intensity and texture, are illustrated in Fig. 15 and 3D in Figure 16. Boundary sharpness (0.62) and homogeneous intensity (0.41) are significant nodes in meningiomas, which encode the "dural tail" sign with boundary gradients >85 HU/mm. Dominant neighbors, including Node 2648, present dural patterns, including homogeneous intensity and unilateral enhancement. In misclassified gliomas, GraphLIME identified feature confounding, where calcification artifacts simulate necrosis, as indicated by high $\beta_{\text{min-intensity}}$ (>0.18). These results demonstrate GraphLIME's capacity to identify diagnostically relevant features and potential sources of errors.

3) Assessment of Significance Ratings: The similarity of node importance scores of GNNExplainer and GraphLIME, as depicted in Figure 17, is also found to be strongly correlated with a Spearman coefficient of 0.79 ($p < 0.001$). While both approaches enjoy consensus on the most important superpixels, GNNExplainer favors structural relations, while GraphLIME emphasizes feature attributes. Areas of disagreement may reflect infiltrative tumor boundaries or artifacts, which are valuable information for model refinement and improving diagnostic accuracy.

VI. Conclusion

In conclusion, the application of GNNExplainer and GraphLIME to our GNN models has elucidated the

decision-making process, highlighting critical subgraphs and features that align with clinical diagnostic criteria. These insights enhance the trustworthiness of our models, facilitating their adoption in clinical settings for brain tumor diagnosis.

References

[1] National Brain Tumor Society, Brain Tumor Facts, 2022, https://braintumor.org/brain-tumors/about-brain-tumors/brain-tumor-facts/.

[2] Johns Hopkins Medicine, Brain Tumor Types, 2021, https://www.hopkinsmedicine.org/health/conditions-and-diseases/brain-tumor/brain-tumor-types.

[3] Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging, PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC10453020/.

[4] CNN-based Graph Neural Network for Brain Tumor Classification, Nature, 2023, https://www.nature.com/articles/s41598-023-41407-8.

[5] A. Ravinder et al., "Enhanced brain tumor classification using graph convolutional neural network architecture," *Scientific Reports*, vol. 13, no. 1, pp. 1–15, 2023.

[6] S. Mishra and D. Verma, "Graph attention autoencoder inspired CNN based brain tumor classification using MRI," *Neurocomputing*, vol. 503, pp. 236–247, 2022.

[7] R. Ying et al., "GNN Explainer: A Tool for Interpreting Graph Neural Networks," *arXiv preprint arXiv:1903.03894*, 2019.

[8] Q. Huang et al., "GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks," *arXiv preprint arXiv:2001.06215*, 2020.

[9] M. Nickparvar, "Brain Tumor MRI Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[11] Y. Wang, J. Li, and Y. Zhang, "Image segmentation of brain MRI based on LTriDP and superpixels of improved SLIC," *BioMed Res. Int.*, vol. 2020, pp. 1–10, 2020.

[12] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[13] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 705–718.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[15] A. Kazi, S. Shekharforoush, A. Krishna, H. Burwinkel, G. Vivar, K. Kortüm, S.-A. Ahmadi, S. Albarqouni, and N. Navab, "InceptionGCN: Receptive field aware graph convolutional network for disease prediction,

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[17] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] M. V. S. da Silva *et al.*, "eXplainable Artificial Intelligence on Medical Images: A Survey," *arXiv preprint arXiv:2305.07511*, 2023.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[21] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Adv. Neural Inf. Process. Syst.*, 2019, pp. 9244–9255.

[22] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "GraphLIME: Local interpretable model explanations for graph neural networks," *arXiv preprint arXiv:2001.06216*, 2020.

[23] H. Yuan, J. Tang, X. Hu, and S. Ji, "XGNN: Towards model-level explanations of graph neural networks," *arXiv preprint arXiv:2006.02587*, 2020.

[24] R. Ani and O. S. Deepa, "An integrated study on convolutional neural networks and graph neural networks for brain tumor classification from MRI images," in *Proc. Int. Conf. Comput. Commun. Intell. Syst.*, 2023, pp. 1–6.