Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 241-250. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1839

Bridging Vision and Graph Attention Learning Using Swin Transformer for Brain Tumor Classification

Radius Tanone[1, 2*], Yos Richard Beeh[2], Pratyaksa Ocsa Nugraha Saian[2], Agus Cahyo Nugroho[3,4]

[1]Artificial Intelligence Research Center, Satya Wacana Christian University, Salatiga, Indonesia
[2]Department of Informatics, Satya Wacana Christian University, Salatiga, Indonesia
[3]Department of Information Management, Chaoyang University of Technology, Taichung City, Taiwan
[4]Department of Information System, Unika Soegijapranata, Semarang, Indonesia

E-mail: radius.tanone@uksw.edu, yosrichardbeeh@uksw.edu, pratyaksa.ocsa@uksw.edu, agus.nugroho@unika.ac.id

Abstract

Accurate brain tumor classification from MRI images is crucial for supporting clinical diagnosis. This study proposes a hybrid framework that integrates a Swin Transformer and Graph-based learning to enhance feature representation and relational modeling. The Swin Transformer, known for its linear computational complexity and multi-scale attention mechanism, serves as a hierarchical feature extractor. The Graph module, on the other hand, captures structural dependencies among the extracted features. The model is evaluated on a four-class brain MRI dataset using 5-Fold cross-validation. Experimental results consistently demonstrate high performance, with a best accuracy of 0.9907. The model achieves sensitivity and specificity of 0.9903 and 0.9970, respectively. Notably, all folds maintain accuracy above 0.9500, indicating strong robustness and generalization capability. The proposed approach effectively reduces misclassification across tumor categories and outperforms baseline methods, underscoring its potential for reliable brain tumor classification.
Keywords: *brain tumor classification, swin transformer, graph transformer, brain tumor, classification.*

1. Introduction

Brain tumors are a global health concern, with the number of cases steadily increasing [1], [2]. Classification is crucial in detecting tumor types based on images obtained from CT scans or MRIs. This classification plays a vital role in supporting clinical decision-making, facilitating early diagnosis, and guiding treatment planning. In the development of artificial intelligence (AI), particularly deep learning [3], [4], which is a part of it, a solution has emerged to aid decision-making. Various models have been developed to achieve this. Conventional computer-aided diagnosis systems have primarily relied on convolutional neural network (CNN) [5], [6], which effectively captures local appearance features but often struggles to model global contextual relationships within complex tumor structures. This is because tumor morphology can vary significantly across patients, necessitating both fine-grained local features and long-range contextual reasoning for accurate recognition, even in MRI images.

Since the advent of attention mechanisms in Natural Language Processing (NLP) [7], Vision Transformer (ViT) [8]has been applied to vision tasks. ViT's ability to capture global context through self-attention enables it to comprehend long-range dependencies in images more effectively than traditional Convolutional Neural Networks (CNN). This makes ViT suitable for medical image analysis. However, the Swin Transformer [9], which is less computationally intensive compared to standard transformers, offers greater efficiency, particularly for high-resolution images. Its hierarchical structure allows it to capture both local and global features effectively, similar to a CNN. Moreover, the Swin Transformer's ability to capture multi-scale representations, including medical images, makes it highly effective. While transformers enhance global context understanding, they lack explicit modeling of structural relationships among image regions.

In contrast to image-based computer vision tasks that rely on convolution or patch-based [10] processes for classification, graphs can be another

viable option for vision tasks. Graph-based deep learning offers a complementary capability by representing image patches or anatomical regions as nodes in a graph, where edges encode meaningful spatial or topological relationships. Graph neural network (GNN) [11], particularly transformer-style graph architectures [12], [13], enable relational reasoning that effectively highlights interactions between different image regions. These characteristics are particularly relevant in brain tumor analysis, especially when analyzing MRI images [14], [15], [16], where understanding the relationships between local tumor textures and the global anatomical context is crucial. Transformer architectures have shown strong performance in vision tasks, but their effectiveness is limited by their reliance on global dependencies. In medical imaging, particularly MRI scans, meaningful diagnostic features often reside in subtle, localized regions that require fine-grained modeling. Traditional graph neural networks, while capable of capturing relational structures, face challenges related to scalability, limited depth, interpretability issues, and sensitivity to noise commonly found in medical images.

Research on brain tumor classification using transformer-based architectures is not a recent development. For example, Tanone et al. [17] applied a Vision Transformer (ViT) for feature extraction, followed by PCA for dimensionality reduction, and CatBoost as the final classifier. Beyond ViT, Swin Transformer–based methods have also been investigated. Bataineh et al. [18] explored a hybrid model incorporating the Swin Transformer and ResNet50V2, while Fu et al. [19] proposed a multimodal framework that integrates Swin Transformer with graph convolution for glioma prognosis. This hybrid design enhances feature representation through cross-modal attention and divergence regularization, yet its performance is still limited by the scarcity of available data. A similar approach was presented by Asiri et al. [20], who utilized the Swin Transformer for brain tumor classification and reported an accuracy of 0.9700. Although the results were promising, the model's effectiveness remained constrained by factors such as a relatively small dataset and suboptimal parameter tuning, preventing it from fully reaching its potential.

Continuing the exploration of Swin Transformer-based models, Pacal [21] proposed an enhanced deep learning framework that incorporates a novel Hybrid Shifted Windows Multi-Head Self-Attention (HSW-MSA) module and a rescaled architecture. In this approach, the traditional MLP component of the Swin Transformer is replaced with a Residual-based MLP (ResMLP), combined with the HSW-MSA mechanism. Despite these architectural innovations, the study provides limited analysis regarding parameter optimization and does not thoroughly evaluate the model's performance across diverse datasets or clinical environments.

Beyond classification tasks, Swin Transformers are widely applied in medical image segmentation. This is illustrated by Hatamizadeh et al. [22], who developed DenseTrans for multimodal segmentation, and by Ranjbarzadeh et al. [23], who introduced an explainable, attention-guided Swin Transformer framework for 3D segmentation. Graph-based methods have also been explored independently for segmentation, such as Ziaeetabar's EfficientGFormer.

Tehsin et al. [24] conducted segmentation experiments using two different datasets, training on one and testing on another. Their study introduced GATransformer, a model equipped with a channel attention module designed to extract richer feature representations from channel–weight interactions. Although the model demonstrated strong performance and efficiency, variations between datasets continue to pose challenges that need to be addressed in future work.

In the field of graph-based modeling, Srivastava et al. [25] conducted a study on the Automated Classification and Grading Diagnosis Model (ACGDM) using MRI images. By utilizing multiple MRI modalities, ACGDM dynamically adjusts its attention across spatial and temporal dimensions, enabling more precise extraction of critical features. However, the model still requires further improvement in hyperparameter optimization, which significantly impacts computational efficiency. Dong et al. [26] proposed a model that integrates the Swin Transformer with a graph-based approach for brain tumor classification. The Swin Transformer captures patch-level features, while a graph attention network models contextual relationships, and the results are fused for final prediction. Although effective, the method still encounters challenges related to image variability, fine-detail representation, and high computational costs.

Recent studies on brain tumor classification have extensively utilized the widely adopted brain tumor dataset consisting of 7023 images. Several approaches have focused on improving baseline deep learning models to achieve higher classification accuracy. For example, Disci et al. [27] employed multiple state-of-the-art pre-trained models, including Xception, MobileNetV2, InceptionV3, ResNet50, VGG16, and DenseNet121, combined with advanced

preprocessing, data augmentation, and transfer learning strategies, achieving an accuracy of 0.9873. Similarly, Kaya et al. [28] proposed Fusion-Brain-Net, a hybrid transfer learning framework integrating VGG16, ResNet50, and MobileNetV2 for enhanced feature extraction through preprocessing, augmentation, feature fusion, fine-tuning, and classification stages, resulting in an accuracy of 0.9708. In addition, Ishaq et al. [29] introduced an improved variant of EfficientNet for multiclass brain tumor detection and classification, achieving an accuracy of 0.9860.

In this research presents several groundbreaking innovations that advance transformer-graph hybrid modeling for brain tumor analysis. First, this research proposes a novel Swin-to-Graph conversion framework that converts transformer-derived feature maps into graph structures, enabling relational reasoning across localized tumor-related image patches. Building on this representation, this research develops a Graph Transformer architecture that incorporates distance-aware and edge-biased attention mechanisms enriched with spatial priors specific to brain tumor morphology. To further enhance predictive capability, this research introduces a multi-scale pooling and classification strategy that integrates mean, max, and sum pooling for more robust graph-level feature aggregation. To clarify the aims and significance of this study, the main contributions are outlined as follows:

1. This study introduce a novel Swin-to-Graph conversion framework that transforms transformer feature maps into graph structures, enabling relational reasoning on brain tumor image patches.
2. This study design a Graph Transformer architecture that incorporates distance-aware and edge-biased attention mechanisms, along with spatial priors from brain tumor images.
3. A comprehensive pooling and classification design is proposed that aggregates multi-scale graph-level features using a combination of mean, max, and sum pooling to enhance the model's performance.

Following the introduction in Section 1, Section 2 outlines the proposed methodology. Section 3 presents the experimental results and their corresponding discussion, and Section 4 concludes the study.

2. Methods

The proposed framework comprises five main stages that collectively transform raw medical images into clinically meaningful predictions. Initially, an image preprocessing phase standardizes input brain MRI scans to enhance feature clarity and ensure consistency across samples. Subsequently, a Swin Transformer Tiny is employed to extract rich hierarchical feature representations from tumor regions. These extracted features are then converted into structured graph representations in the graph construction stage, where patch relationships and spatial dependencies are encoded.

The fourth stage involves Graph Transformer modeling, which leverages distance-aware and edge-biased attention to reason over graph-connected patches and refine feature interactions. Finally, a classification module aggregates graph-level information to generate the final tumor classification predictions. Figure 1 visually summarizes the complete two-step workflow of these stages.

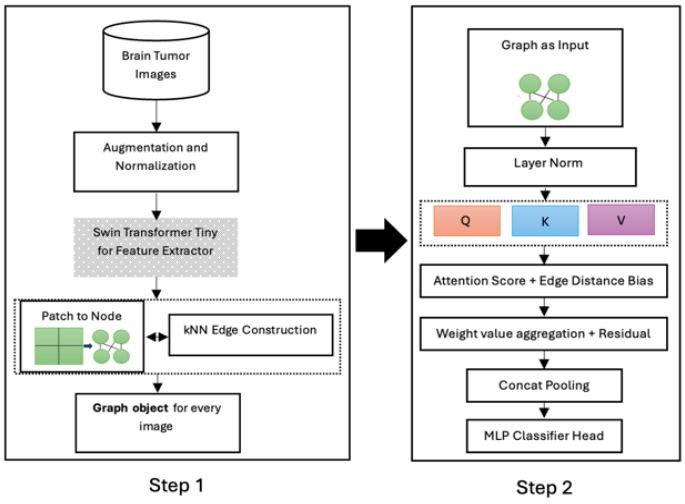
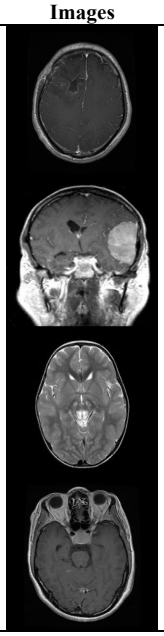


Figure 1. Proposed method.

2. 1 Dataset and Preprocessing

The brain tumor dataset employed in this study consists of MRI images obtained from a publicly available Kaggle repository [30]. The use of a public dataset was considered appropriate because primary medical brain tumor imaging datasets are difficult to acquire due to privacy, ethical, and accessibility constraints. Furthermore, this dataset has been widely utilized in many previous brain tumor classification studies, enabling fair comparison and benchmarking with existing methods. In total, the dataset contains 7,023 image samples representing different tumor categories, providing sufficient variability for model development, training, and evaluation. A detailed breakdown of the dataset composition and class distribution is presented in Table 1.

Table 1. Dataset description.

Classes	Images	Total
Glioma		1622
Meningioma		1646
No Tumor		2001
Pituitary		1758

According to Table 1, this study utilizes a publicly accessible MRI brain tumor dataset (gray scale) comprising four categories: glioma, meningioma, pituitary tumor, and no tumor. To ensure consistency and enhance model performance, each image undergoes several preprocessing stages. Initially, all images are resized to 224 × 224 pixels. Subsequently, pixel values are normalized to the range of [0,1]. Furthermore, data augmentation techniques, such as random rotation, flipping, and intensity perturbation, are employed during training to bolster model robustness and generalization capabilities.

2.2 Feature Extraction using Swin Transformer

In this research, a pretrained Swin Transformer Tiny (Swin-T) model is used as the feature extractor. The Tiny Swin Transformer pretrained model was selected due to several strengths. It offers linear computational complexity and a hierarchical feature structure capable of generating multi-scale feature maps, much like traditional CNNs. Its "shifted window" strategy allows information sharing between adjacent regions across layers. In addition to achieving state-of-the-art performance, Swin-T's window-based design is efficient in memory usage and well suited for hardware implementation. Moreover, using Swin Transformer as a feature extractor is especially beneficial, as its localized self-attention effectively highlights and preserves details in weakly textured or complex image areas—conditions where many other methods tend to struggle. Furthermore, the classification head is removed, retaining only the hierarchical encoder. Each image is passed through the Swin backbone to generate a spatial feature map as shown in Eq (1). The proposed framework converts an input MRI image Iinto a structured representation for classification. First, each image is resized and normalized, yielding a preprocessed form $\hat{I}$. The normalized image is then passed through a Swin Transformer encoder, producing a latent feature collection as shown in equation (1).

$$F = \text{Swin}(\hat{I}) = \{f_i \mid f_i \in \mathbb{R}^d\} \quad (1)$$

where $f_i$denotes the feature representation extracted from local patches of the image. These patch-level embeddings are interpreted as graph nodes, forming a node set V.

2.3 Image-to-Graph Construction

To model spatial interactions, this research construct a graph $G = (V, E)$, where edges are defined through k-nearest neighbor relationships among patches. For each node v_i, its neighborhood $\mathcal{N}_k(i)$is determined by the k closest patches, resulting in edge connections (i, j)stored in the set E. The constructed graph is processed using a Graph Transformer consisting of L attention layers. Node features are initialized via a linear projection as shown in equation (2):

$$X^{(0)} = \phi(V) \quad (2)$$

and then iteratively updated through message passing. At each layer ℓ, pairwise attention coefficients are computed as shown in equation (3):

$$\alpha_{ij} = \frac{(W_Q x_i^{(\ell)})^\top (W_K x_j^{(\ell)})}{\sqrt{d_h}} + b_{ij}, \quad (3)$$

where W_Q, W_K, and W_V are learnable projection matrices, d_h refers to head dimension and b_{ij} is an edge-aware bias term. The normalized attention weights are obtained in equation (4):

$$a_{ij} = \text{softmax}_j(\alpha_{ij}) \quad (4)$$

and aggregated messages for node i are computed in equation (5):

$$m_i = \sum_{j \in \mathcal{N}_k(i)} a_{ij} \cdot (W_V x_j^{(\ell)}) \quad (5)$$

The updated node representation is then shown in equation (6):

$$x_i^{(\ell+1)} = \text{FFN}(\text{Norm}(x_i^{(\ell)} + m_i)) \quad (6)$$

where FFN refers to a feed-forward network with residual connections and normalization. After L propagation layers, the graph representation z is formed through a pooling operation combining mean, max, and summation statistics as shown in equation (7):

$$z = \text{Concat}(\text{Mean}(X^{(L)}), \text{Max}(X^{(L)}), \text{Sum}(X^{(L)})) \quad (7)$$

Next, the pooled representation is passed through a classification network defined in equation (8):

$$y = W_3\, \sigma(W_2\, \sigma(W_1 z + b_1) + b_2) + b_3 \quad (8)$$

followed by softmax activation to produce per-class probabilities $\hat{y} = \text{Softmax}(y)$, and finally, the predicted class label is computed as $\hat{c} = \arg\max(\hat{y})$. The pseudocode of the network can be seen in Figure 2.

```
ALGORITHM BrainTumorClassification(images,
                    labels):
        Step 1: Convert images to graphs
                    graphs = []
            FOR each image in images:
features = SwinTransformer(image)    // 768-dim
patches = DivideIntoPatches(features)//Multiple
                    nodes
  edges = ConnectKNN(patches, k=8)//kNN graph
   graph = Graph(nodes=patches, edges=edges)
              graphs.append(graph)
          Step 2: Build and train model
            model = GraphTransformer(
        layers=3,heads=8,hidden_dim=512)
             Step 3: K-fold training
              FOR fold in 1 to 5:
 train_graphs, val_graphs = SplitData(graphs,
                    fold)

             FOR epoch in 1 to 100:
           FOR batch in train_graphs:
          pred = model.forward(batch)
    loss = CrossEntropy(pred, batch.labels)
                loss.backward()
               optimizer.step()
    accuracy = Evaluate(model, val_graphs)
    results[fold] = CalculateMetrics(model,
                    val_graphs)
            Step 4: Report results
      mean_acc = Mean(results.accuracies)
     PRINT "Mean Accuracy: {mean_acc:.4f}"

            RETURN model, results
              END ALGORITHM
```

Figure 2. The pseudocode of proposed network.

Table 2 summarizes the hyperparameters used in this study. These parameters were selected to ensure stable training and optimal performance of the proposed model, including k_neighbors, distance_metric, and other key training settings, which were applied consistently across all experiments.

Table 2. Hyperparameters setting.

Hyperparameter	Value	Description
k neighbors	8	Number of nearest neighbors for kNN graph construction
distance metric	euclidean	Distance metric for kNN graph
number of layers	3	Number of Graph Transformer layers
number of heads	8	Number of attention heads per layer
edge features	True	Use edge features (spatial distance)
hidden dim	512	Hidden dimension size
head dim	64	Dimension per attention head (512/8)
ffn hidden dim	2048	Feed-forward network hidden dimension (512×4)
dropout	0.3	Dropout rate
pooling method	concat	Graph pooling method (mean + max + sum)
pooled dim	1536	Pooled feature dimension (512×3)

2.5 Evaluation and Visualization

To assess the performance of the proposed Swin Transformer and Graph-based approach, multiple evaluation metrics [31] were applied. Accuracy served as the main indicator, reflecting the percentage of brain tumors correctly classified. To gain deeper insight into class-level performance and address dataset imbalance, precision, recall, and F1-score were also computed. Precision reflects how reliable the positive predictions are, recall indicates the model's ability to detect tumor cases, and the F1-score provides a balanced measure between the two. Furthermore, confusion matrices were used to illustrate patterns of misclassification across different tumor categories. Collectively, these metrics offer a thorough evaluation of the model's robustness and its improvements compared to baseline approaches. Let TP as True Positive, TN as True Negative, FP as False Positive and FN = False Negative, the mathematical definitions of these metrics are provided in Eq (9–14) below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (11)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (13)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (14)$$

3. Results and Discussion

3.1 Environmental Setting

All experiments were conducted on a MacBook Air equipped with an Apple M1 chip, featuring an 8-core CPU/GPU architecture and 16 GB of unified memory. The development environment was set up using the Anaconda Python distribution to ensure consistent and reproducible package management. Model training and evaluation were performed using Jupyter Notebook.

PyTorch (ARM-compatible version) was the primary deep learning framework used for model development. Other libraries utilized in the project include NumPy, Pandas, Matplotlib, scikit-learn, and torchvision. To optimize computational efficiency, the environment was optimized using an Apple Silicon-compatible PyTorch backend.

3.2 Results

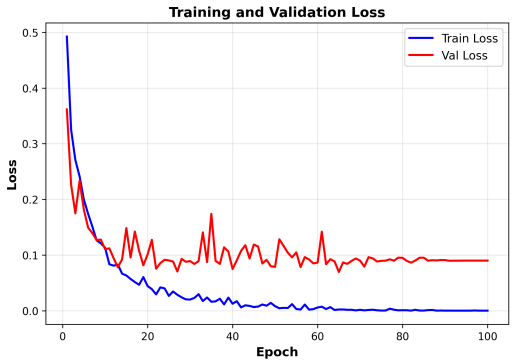
This section presents the model's performance metrics, along with training and validation curves, and the resulting confusion matrix. An ablation study is also included to evaluate how each component contributes to the model's overall effectiveness.

Table 3. Best Result of the model (Fold 5).

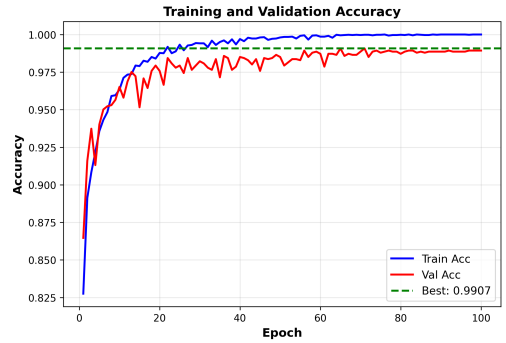
Class	Sens	Spec	Prec	Acc
glioma	0.9846	0.9981	0.9938	0.9947
meningioma	0.9878	0.9926	0.9760	0.9914
no tumor	0.9975	0.9990	0.9975	0.9986
pituitary	0.9915	0.9981	0.9943	0.9961
Average	**0.9903**	**0.9970**	**0.9904**	**0.9907**

Table 3 reveals that Fold 5 demonstrated the most impressive performance during K-Fold cross-validation. The model achieved an accuracy of 0.9907. While accuracy alone doesn't fully encapsulate performance, the average sensitivity of 0.9903 and specificity of 0.9970 further underscore the model's exceptional ability to distinguish tumor types. These findings suggest

that, with a relatively balanced dataset, the model maintains consistent classification performance. The training and loss curves corresponding to Fold 5 are presented in Figure 3 (a) and (b).



(a)



(b)

Figure 3. (a) Training, and (b) validation curve of Fold 5.

Additionally, the accuracy results for the remaining folds are illustrated in Figure 4. Among them, Fold 3 recorded the lowest accuracy at 0.9758, while Folds 1, 2, and 4 show comparable performance. These findings indicate that, even across different cross-validation splits, the model consistently achieves an accuracy above 0.9500.

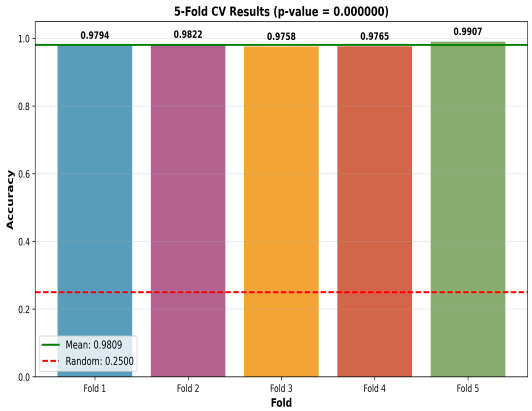


Figure 4. Performance of cross validation.

Figure 5 presents the confusion matrix for Fold 5. The glioma class shows the highest number of misclassifications, with 5 errors and 319 correct predictions. In contrast, the no-tumor class records the fewest errors, with only one misclassified case. The remaining classes also exhibit low error counts, with 4 misclassifications for meningioma and 3 for pituitary.

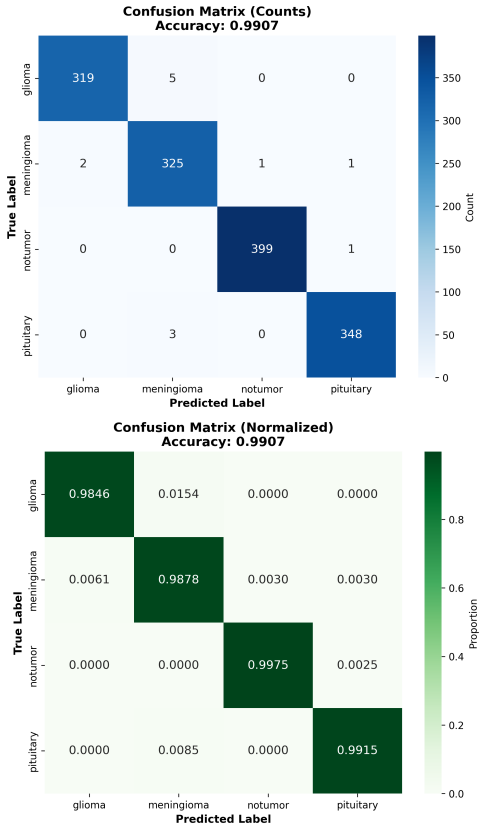


Figure 5. Confusion matrix of Fold 5.

3.3 Ablation study

To see the effectiveness of the modules in the proposed method, this research provide an ablation study. Table 4 presents crucial ablation results showing how the model affects the removal of components of the proposed method.

Table 4. Ablation study summary.

Configuration	Acc	Sens	Spec
1 Layer	**0.9851**	0.9755	0.9921
2 Layers	0.9843	**0.9819**	**0.9941**
Full Model	0.9822	0.9781	0.9930
Without Edge Features	0.9822	0.9726	0.9913
Max Pooling Only	0.9822	0.9717	0.9909
Mean Pooling Only	0.9815	0.9690	0.9900
4 Attention Heads	0.9808	0.9782	0.9929

While our proposed method uses multiple layers, while the 1-layer configuration achieves slightly higher accuracy (0.9851), the full 3-layer model demonstrates superior balance across

sensitivity, and specificity, confirming that stable layer depth contributes to the generalization of model performance. Furthermore, the removal of edge features leads to a consistent decrease in sensitivity (-0.0055), validating their role in spatial representation, which is crucial for understanding image data. Furthermore, replacing concat with either average-only or maximum-only readout decreases all metrics, indicating that the concat strategy in our proposed method captures a more comprehensive graph-level representation. Ablation study also included reducing the number of attention heads from 8 to 4, resulting in minimal accuracy changes but a slight decrease in sensitivity, suggesting that the number of heads plays a secondary role relative to the concat strategy and edge features. So, from the conclusion that can be drawn, our proposed method has a number of layers, edge features, pooling strategy and number of heads that affect the model performance.

3.4 Discussion

The experimental results demonstrate that the proposed model effectively integrates Swin Transformer and Graph-based learning for brain tumor classification. The K-Fold cross-validation evaluation provides strong evidence of the model's stability and generalization ability across different data splits. Among all folds, Fold 5 achieved the highest performance, with an accuracy of 0.9907, sensitivity of 0.9903, and specificity of 0.9970. These values indicate that the model is highly capable of identifying tumor regions while minimizing false predictions in non-tumor cases. Such balanced results are crucial in medical imaging tasks, where both false positives and false negatives must be minimized for reliable clinical decision-making.

Although Fold 5 exhibited the best performance, the other folds also consistently produced high accuracy levels. Even the lowest-performing fold (Fold 3) achieved an accuracy of 0.9758, demonstrating the model's robustness under varying data distributions. The consistency across Folds 1, 2, and 5, with accuracies above 0.9800, further suggests that the model's architecture successfully captures discriminative tumor features despite potential variance in sample composition. Overall, maintaining accuracy above 0.9500 across all folds highlights the reliability of the proposed approach for real-world deployment.

The confusion matrix analysis provides deeper insights into the class-level behavior of the model. The glioma class exhibited the highest number of misclassifications, with five incorrect predictions. This could be attributed to the structural similarity between glioma and other tumor types, as well as the inherent variability in glioma appearance within MRI scans. On the other hand, the no-tumor class showed extremely high classification accuracy, with only one misclassified instance, suggesting that the model distinguishes normal cases with strong confidence. Misclassifications in the meningioma and pituitary classes remained low (four and three cases, respectively), indicating that the model's feature representations generalize well across tumor categories.

Another key factor contributing to the model's effectiveness is the Swin Transformer's hierarchical feature extraction and localized attention mechanism. These properties enable the network to capture multi-scale contextual information, which is crucial for distinguishing subtle differences between tumor types. The Graph module further enhances relational reasoning between extracted features, improving the model's ability to recognize spatial and structural patterns in complex MRI data. Taken together, these components create a powerful, complementary architecture capable of handling both global and local dependencies. A comparative analysis with previous research was conducted to assess the effectiveness of the proposed model.

To compare the effectiveness of our proposed method, this research conducted state-of-the-art experiments to assess the performance of the baseline model and other model variants. Table 5 shows the results of experiments with hyperparameters such as an image size of 64x64, 20 epochs, and no cross-validation. Although the Swin-T (baseline) is higher than other models, this demonstrates that the Graph, the foundation of our proposed method, is quite influential in improving model performance.

Table 5. State-of-the-art performances.

Model	Acc	Sens	Spec
Swin-T (baseline)	0.8747	0.8679	0.9584
Swin-B	0.8733	0.8670	0.9579
Swin-S	0.8598	0.8529	0.9531
ConvNeXt-T	0.8349	0.8275	0.9447
Swin-T + Graph Attention	**0.9907**	**0.9903**	**0.9970**

Despite the excellent results, some limitations remain. The presence of misclassifications, particularly in the glioma class, suggests that additional strategies, such as stronger data augmentation, class rebalancing, or the incorporation of multimodal imaging (e.g., T1, T2, FLAIR), may further improve performance. Another potential direction is leveraging medical-

domain pretrained models to reduce sensitivity to variability in tumor shape and texture. Overall, the findings highlight that the combination of the Swin Transformer and Graph-based learning provides a robust, high-performing solution for brain tumor classification. This approach outperforms other models such as Swin-T (baseline), Swin-B, Swin-S, and ConvNeXt-T on metrics including accuracy, sensitivity, and specificity. The strong cross-validation results, minimal misclassification rates, and consistent metric values across folds demonstrate the proposed model's potential for practical clinical applications.

4. Conclusion

This study presented a brain tumor classification framework that combines the strengths of the Swin Transformer and Graph-based learning to effectively extract and model multi-scale structural information from MRI images. Experimental results across 5-Fold cross-validation demonstrate that the proposed model delivers highly consistent and reliable performance, with Fold 5 achieving the best accuracy of 0.9907 and overall results remaining above 0.9500 across all folds. The sensitivity and specificity values further confirm the model's capability to correctly identify tumor cases while minimizing false detections.

Analysis of the confusion matrix shows that the model performs strongly across all tumor categories, with only a small number of misclassifications, particularly in the glioma class due to its high visual variability. The integration of hierarchical attention from the Swin Transformer and relational reasoning from the Graph module significantly contributes to the model's effectiveness in capturing both local and global tumor features. Overall, the study demonstrates that the proposed method is a robust and high-performing solution for brain tumor classification. With its strong generalization ability and low misclassification rates, the framework holds promise for deployment in real-world clinical decision support systems. Future improvements may include expanding the dataset, incorporating multi-modal MRI sequences, and applying domain-specific pretraining to further enhance classification accuracy.

Acknowledgements

The authors would like to express their sincere gratitude to Universitas Kristen Satya Wacana for the financial support provided through Research Grant Contract No. 203/SPK-PF/RIK/08/2025.

References

[1] N. Soni *et al.*, "Meningioma: Molecular Updates from the 2021 World Health Organization Classification of CNS Tumors and Imaging Correlates," *Am. J. Neuroradiol.*, vol. 46, no. 2, pp. 240–250, Feb. 2025, https://doi.org/10.3174/AJNR.A8368.

[2] S. Taslimi, S. Brogly, T. P. Hanna, J. Shellenberger, M. Whitehead, and R. Alkins, "A Population-Based Cohort Study of Glioblastoma (World Health Organization Grade 4 Gliomas) in Ontario: Continued Improvement in Care over 25 Years," *World Neurosurg.*, vol. 196, p. 123821, Apr. 2025, https://doi.org/10.1016/J.WNEU.2025.123821.

[3] G. Jayaraman, S. Meganathan, S. S. M. Shah, M. Anuradha, R. K. Sundararajan, and R. Rajakumar, "A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification," *Sci. Reports 2025*, Nov. 2025, https://doi.org/10.1038/s41598-025-28636-9.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, https://doi.org/10.1038/nature14539.

[5] Z. Rasheed *et al.*, "Brain Tumor Classification from MRI Using Image Enhancement and Convolutional Neural Network Techniques," *Brain Sci.*, vol. 13, no. 9, p. 1320, Sep. 2023, https://doi.org/10.3390/brainsci13091320.

[6] S. Saifullah, A. P. Suryotomo, R. Drężewski, R. Tanone, and T. Tundo, "Optimizing Brain Tumor Segmentation Through CNN U-Net with CLAHE-HE Image Enhancement," in *1st Int. Conf. Adv. Informatics Intell. Inf. Syst. (ICAI3S 2023), 2023*, 2024, pp. 90–101. https://doi.org/10.2991/978-94-6463-366-5_9.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762.

[8] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, https://doi.org/10.48550/arxiv.2010.11929.

[9] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9992–10002, Mar. 2021, https://doi.org/10.48550/arxiv.2103.14030.

[10] A. Trockman and J. Z. Kolter, "Patches Are All You Need?," 2022, Accessed: Mar. 29, 2022. [Online]. Available: https://github.com/locuslab/convmixer.

[11] J. Zhou *et al.*, "Graph Neural Networks: A Review of Methods and Applications," *AI Open*, vol. 1, pp. 57–81, Dec. 2018, https://doi.org/10.1016/j.aiopen.2021.01.001.

[12] A. Shehzad, F. Xia, S. Abid, C. Peng, S. Yu, D. Zhang, and K. Verspoor, "Graph Transformers: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, early access, Jan. 2026, https://doi.org/10.1109/TNNLS.2025.3646122.

[13] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph Transformer Networks," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 11960–11970. https://doi.org/10.48550/arXiv.1911.06455.

[14] I. S. Rajput, A. Gupta, V. Jain, and S. Tyagi, "A transfer learning-based brain tumor classification using magnetic resonance images," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 20487–20506, Mar. 2024, https://doi.org/10.1007/s11042-023-16143-w.

[15] R. Javed, T. Abbas, A. H. Khan, A. Daud, A. Bukhari, and R. Alharbey, "Deep learning for lungs cancer detection: a review," *Artif. Intell. Rev.*, vol. 57, no. 8, p. 197, Aug. 2024, https://doi.org/10.1007/s10462-024-10807-1.

[16] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, "Vision Transformers, Ensemble Model, and Transfer Learning Leveraging Explainable AI for Brain Tumor Detection and Classification," *IEEE J. Biomed. Heal. Informatics*, vol. 28, no. 3, pp. 1261–1272, Mar. 2024, https://doi.org/10.1109/JBHI.2023.3266614.

[17] R. Tanone, L.-H. Li, and S. Saifullah, "ViT-CB: Integrating hybrid Vision Transformer and CatBoost to enhanced brain tumor detection with SHAP," *Biomed. Signal Process. Control*, vol. 100, p. 107027, 2025, https://doi.org/10.1016/j.bspc.2024.107027.

[18] A. F. Al Bataineh *et al.*, "Enhanced Magnetic Resonance Imaging-Based Brain Tumor Classification with a Hybrid Swin Transformer and ResNet50V2 Model," *Appl. Sci. 2024, Vol. 14, Page 10154*, vol. 14, no. 22, p. 10154, Nov. 2024, https://doi.org/10.3390/APP142210154.

[19] M. Fu, M. Fang, R. A. Khan, B. Liao, Z. Hu, and F. X. Wu, "SG-Fusion: A swin-transformer and graph convolution-based multi-modal deep neural network for glioma prognosis," *Artif. Intell. Med.*, vol. 157, p. 102972, Nov. 2024, https://doi.org/10.1016/J.ARTMED.2024.102972.

[20] A. A. Asiri *et al.*, "Advancing brain tumor detection: harnessing the Swin Transformer's power for accurate classification and performance analysis," *PeerJ Comput. Sci.*, vol. 10, p. e1867, Feb. 2024, https://doi.org/10.7717/PEERJ-CS.1867.

[21] I. Pacal, "A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images," *Int. J. Mach. Learn. Cybern. 2024 159*, vol. 15, no. 9, pp. 3579–3597, Mar. 2024, https://doi.org/10.1007/S13042-024-02110-W.

[22] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12962 LNCS, pp. 272–284, 2022, https://doi.org/10.1007/978-3-031-08999-2_22.

[23] R. Ranjbarzadeh, A. Keles, S. Anari, P. Zarbakhsh, and M. Bendechache, "Explainable Attention-Guided Swin Transformer Networks for Brain Tumor Segmentation from 3D MRI," pp. 110–128, 2025, https://doi.org/10.1007/978-3-032-11512-6_11.

[24] S. Tehsin, I. M. Nasir, and R. Damaševičius, "GATransformer: A Graph Attention Network-Based Transformer Model to Generate Explainable Attentions for Brain Tumor Detection," *Algorithms 2025, Vol. 18, Page 89*, vol. 18, no. 2, p. 89, Feb. 2025, https://doi.org/10.3390/A18020089.

[25] S. Srivastava, P. Jain, S. K. Pandey, G. Dubey, and N. N. Das, "Automated Brain Tumor Classification and Grading Using Multi-scale Graph Neural Network with Spatio-Temporal Transformer Attention Through MRI Scans," *Interdiscip. Sci. Comput. Life Sci. 2025*, pp. 1–29, Jun. 2025, https://doi.org/10.1007/S12539-025-00718-2.

[26] C. Dong, D. Sun, and B. Luo, "HLAE: Hierarchical Local Attention Encoder for MRI Brain Tumor Image Classification," *Cogn. Comput. 2025 171*, vol. 17, no. 1, pp. 54-, Jan. 2025, https://doi.org/10.1007/S12559-025-10419-Z.

[27] R. Disci, F. Gurcan, and A. Soylu, "Advanced Brain Tumor Classification in MR Images Using Transfer Learning and Pre-Trained Deep CNN Models," *Cancers 2025, Vol. 17, Page 121*, vol. 17, no. 1, p. 121, Jan. 2025, https://doi.org/10.3390/CANCERS17010121.

[28] Y. Kaya, E. Akat, and S. Yıldırım, "Fusion-Brain-Net: A Novel Deep Fusion Model for Brain Tumor Classification," *Brain Behav.*, vol. 15, no. 5, p. e70520, May 2025, https://doi.org/10.1002/brb3.70520.

[29] A. Ishaq, F. U. M. Ullah, P. Hamandawana, D. J. Cho, and T. S. Chung, "Improved EfficientNet Architecture for Multi-Grade Brain Tumor Detection," *Electron. 2025, Vol. 14, Page 710*, vol. 14, no. 4, p. 710, Feb. 2025, https://doi.org/10.3390/ELECTRONICS14040710.

[30] Nickparvar M., "Brain Tumor MRI Dataset." Accessed: May 03, 2024. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

[31] P. Christen, D. J. Hand, and N. Kirielle, "A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives," Oct. 06, 2024, *ACMPUB27New York, NY.*, https://doi.org/10.1145/3606367.