CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Dissecting Glioma Heterogeneity: A Deep Hybrid Graph Convolutional Network With Hinge Attention for Causal-Effect Explainability

Sultanul Arifeen Hamim[1] | Mubasshar U. I. Tamim[1] | S. M. Hasan Mahmud[2] | M. F. Mridha[1] | Zeyar Aung[3]

[1]Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh | [2]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh | [3]Department of Computer Science, Khalifa University, Abu Dhabi, UAE

Correspondence: M. F. Mridha (firoz.mridha@aiub.edu) | Zeyar Aung (zeyar.aung@ku.ac.ae)

Received: 4 November 2025 | **Revised:** 24 March 2026 | **Accepted:** 7 May 2026

Keywords: brain tumour classification | convolutional neural networks (CNNs) | explainable AI (XAI) | graph neural networks (GNNs) | hybrid deep learning model

ABSTRACT

Brain tumour classification is a critical task in medical imaging that requires accurate and interpretable solutions to assist in clinical decision-making. In this paper, we present GraphConvNet-X, a novel hybrid model that integrates convolutional neural networks (CNNs) for spatial feature extraction with graph neural networks (GNNs) that employ a novel hinge attention network (HAN) for robust relational aggregation. Leveraging Explainable Artificial Intelligence (XAI) techniques, such as gradient-weighted class activation mapping (Grad-CAM) and SHAP, the model provides transparent insights into its predictions, aligning them with clinically significant regions in magnetic resonance imaging (MRI) scans. The integration of HAN enhances the GNN's ability to selectively focus on the most salient tumour-related graph features, improving overall discriminative power. The proposed architecture was evaluated on three benchmark datasets, achieving impressive accuracies of 98%, 97% and 96%, surpassing state-of-the-art methods. Extensive experiments have demonstrated its robustness across various configurations and highlighted its adaptability to diverse scenarios. By combining precision, adaptability and interpretability, GraphConvNet-X sets a new standard for AI-driven diagnostic tools, showing its potential to revolutionise brain tumour classification and contribute significantly to modern healthcare.

1 | Introduction

Brain tumours are among the most serious and potentially life-threatening health ailments, posing serious challenges for their diagnosis and treatment. Early diagnosis of brain tumours is crucial for improving patient outcomes, as a delay in detection can lead to serious complications or even death. A wide range of tumour types constitute brain tumours, such as gliomas, meningiomas and pituitary tumours, each of which requires different treatment strategies [1, 2]. The identification of different tumour groups, in addition to their separation from non-tumourous lesions of the brain, is a complex process that requires high accuracy. Magnetic Resonance Imaging (MRI) is a common imaging technique used in hospitals for the diagnosis of brain tumours [3, 4]. MRI provides high-resolution structural and spatial details of brain tissues, thereby aiding in diagnosis by radiologists. Interpretation of MRI is a time-consuming process that is subjective in nature and intersubject sensitive. The complexity of MRI images, in conjunction with minor differences between tumour groups, is responsible for potential inaccuracies in diagnosis. As a result, there is a high demand for

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

computer-based and accurate diagnostic systems to diagnose brain tumours accurately [5, 6].

Recent advancements in artificial intelligence (AI) and machine learning (ML) have significantly facilitated medical image analysis [7]. The application of convolutional neural networks (CNNs) is a common practice in image-based classification because of their ability to extract spatial features [8, 9]. Models developed using CNNs have been found to be highly accurate in a range of applications in medical imaging, including tumour classification. However, CNNs predominantly extract spatial features without modelling relationships between different parts of a structure in an explicit manner, which would be helpful in interpreting complex tumour structures [10, 11]. Graph Neural Networks (GNNs) better capture relational and topological details and are more helpful in modelling interactions between different parts of a structure. However, GNNs are incapable of modelling sophisticated spatial patterns compared with CNNs. With such a trade-off in mind, a combined approach that leverages the strength of GNNs and CNNs is required to improve the precision of brain tumour classification [12, 13]. Another crucial aspect of AI-based diagnosis is interpretability. Deep learning models of the nature of CNNs and GNNs are typically black-box systems, which provide little explanation of their decision-making processes. Lack of interpretability is a challenge for medical practitioners because transparency is a prerequisite to be trusted and to cross-check AI-based diagnoses. XAI methods mitigate such a challenge to a great extent by making AI models transparent, understandable and usable in a clinical setup [14, 15].

To address these challenges, in this study, GraphConvNet-X is introduced, a new hybrid deep learning method that combines CNNs and an advanced GNN, utilising a novel Hinge Attention Network (HAN) for superior feature aggregation, to classify brain tumours. By adopting XAI methods, the proposed method attains high classification precision and provides understandable interpretations in its prediction process, making it more useful in actual clinical use cases. The proposed method was compared with three publicly released brain tumour MRI datasets to provide robustness and generalisability. Furthermore, a comparative analysis of the basic deep learning method revealed the higher precision, interpretability and usability of the proposed hybrid structure in a clinical setup.

The key contributions of this work are as follows:

1. This study presents GraphConvNet-X, a novel hybrid CNN–GNN architecture that integrates spatial feature extraction with relational reasoning for improved brain tumour classification.
2. We introduce a Hinge Attention Network (HAN), a margin-based attention mechanism that selectively suppresses weak feature interactions, enhancing robustness and interpretability in graph learning.
3. We develop an explainable AI framework combining Grad-CAM and SHAP to provide both spatial and relational interpretability aligned with clinically relevant regions.
4. We demonstrate the effectiveness and generalisability of the proposed model across three benchmark MRI datasets, achieving superior performance compared to state-of-the-art CNN and transformer-based models.

The remainder of this paper is organised as follows. Section 2 provides a detailed review of related work and highlights the limitations of the existing CNN and GNN models. Section 3 describes the methodology, including the architecture of GraphConvNet-X and its integration with explainable AI techniques. Section 4 presents the results, showcasing the performance through a quantitative analysis, qualitative visualisations and ablation studies. The Section 5 discuss the paper. Finally, Section 6 concludes the study by summarising the key findings and discussing future research directions.

2 | Literature Review

The classification of brain tumours using MRI scans has been a dynamic line of research driven by the need for accurate and timely diagnosis to treat patients more effectively. Classical machine learning methods provide initial solutions, but newer developments in deep learning, such as CNNs, profoundly transform the results of classification. Furthermore, models that combine relational learning using GNNs with interpretability using XAI methods have become increasingly common. This section discusses the existing literature on brain tumour classification, pointing towards advanced hybrid models and presents the gaps addressed in the proposed GraphConvNet-X model.

Traditional machine learning approaches have been employed in current work in computerised system design for automatic tumour classification of the brain. Support vector machines (SVM), decision trees, naïve Bayes, K-nearest neighbours (KNN) and linear discriminant analysis (LDA) approaches have been employed extensively to classify brain tumours using MRI images. Consequently, SVM and KNN classifiers based on the texture and statistical properties of brain MRI images attained a precision of 85% and 82%, respectively, on a set of 1500 samples [7]. Similarly, in a recent study carried out by Arshad Choudhry et al., a hybrid GNN–CNN framework incorporating AL-CAM, contrastive learning, and excitation backpropagation was proposed for interpretable brain-tumor segmentation, achieving 95% accuracy and an AUROC of 0.96, with superior Dice, IoU, and HD95 performance [5]. Despite such efforts, traditional approaches lack in that their handcrafted nature is time-consuming to design and is not capable of modelling underlying sophisticated patterns in medical imagery. These approaches also exhibit high sensitivity to imaging condition changes and dataset quality, resulting in different results when employed with different datasets. For example, when employed in a different set of samples, the SVM classifier resulted in a precision of 78%, which is a failure of traditional approaches to be transferable to different environments.

CNNs have become widely used in medical imaging applications because of their capacity to learn hierarchical spatial features. Abiwinanda et al. [8] introduced a basic CNN structure with one convolutional layer that could obtain a precision of 84.19% using a dataset of 3064 MRI images. Pashaei et al. [12] also introduced a five-layered CNN structure that was able to obtain a higher precision of 93.68% using more deeply stacked feature extraction layers. Nayak et al. also investigated different pre-trained CNN architectures such as MobileNet, ResNet-50 and EfficientNet, and obtained a maximum precision of 95%

using MobileNet [13]. The results indicate that CNNs are efficient for classifying brain tumours but also indicate their weakness in interpretability and relational knowledge. Hybrid CNN models have also been investigated to improve their performance. Sajid and Hussain [16] introduced a two-path and three-path network-based hybrid CNN model that was able to obtain a precision of 86%. However, the high complexity of the hybrid models results in poor interpretability. Hashemzehi et al. [17] introduced a CNN-Neural Autoregressive Distribution Estimation (NADE)-based hybrid model that could obtain a precision of 95% and indicated that different deep learning strategies can be combined to achieve high precision.

Hybrid architectures that combine GNNs have been proposed to address the limitations of spatial-only learning in CNNs. GNNs better capture relational dependencies and hold promise in enhancing medical-imaging applications. Ravinder et al. [1] proposed a hybrid method that combined a Graph Convolutional Network (GCN) with a 26-layer CNN to enhance the accuracy of brain tumour classification. Using a dataset of 3094 MRI images, this study explored various adjacency matrix methods and achieved a maximum precision of 95.01% using the best configuration. This work indicates that there is a need for non-Euclidean representations of imaging data to capture relational dependencies to enable better diagnosis. Similarly, Arshad Choudhry et al. [5] proposed a multimodal MRI-based method for brain tumour segmentation that combined GNNs and CNNs. The work achieved a Dice score of 97.43% using explainability methods, such as Adaptive Learning Class Activation Maps (AL-CAM) and contrastive Excitation Backpropagation (c-EB). This method indicates that hybrid models can provide high precision in conjunction with high interpretability without sacrificing computational efficiency. Another study by Ashok et al. [18] investigated the use of GCNs in conjunction with pre-trained CNNs, ResNet18, VGG16 and DenseNet121, for brain tumour classification using MRI. The results showed that classification accuracies between 64% and 74% were noted, with ResNet18 being the most accurate. Although this study utilised sophisticated graph-based representations and Principal Component Analysis (PCA) to boost feature learning, it also emphasised that optimisation and scalability to large datasets need to be explored. Finally, a combined model using GNNs and CNNs was found to be up to 93.68% accurate in brain tumour classification, indicating that combined architectures hold the potential to utilise both relational and spatial knowledge [3]. These studies indicate that there is a trend towards using combined CNNs and GNNs to facilitate more accurate classifications.

Recent studies have demonstrated the effectiveness of attention-driven hierarchical representation learning in complex visual reasoning tasks. For example, Tang et al. [19] introduce a region-to-point localisation framework that leverages hierarchical semantic representations to improve robustness and spatial reasoning. Similarly, Li et al. [20] propose a vision-only dynamic Gaussian formulation for realistic driving simulation, highlighting the importance of structured and interpretable feature modelling. Furthermore, recent works on unsupervised pre-training with language–vision prompts have shown that multimodal attention mechanisms can significantly improve generalisation in low-data regimes [21]. These methods represent a broader trend towards attention-based relational modelling that moves beyond traditional convolutional feature extraction techniques. Following this line of research, GraphConvNet-X employs a graph-centric approach with a hinge-based attention mechanism to directly model the relationships between latent feature groupings. Unlike traditional softmax-based attention methods, the proposed hinge-based gating function suppresses weak relationships to produce more robust relational reasoning.

Explainability plays a significant role in such cases because it helps build trust and makes the tool more user-friendly. Most of the CNN-based approaches lack explainability, and this makes it difficult to use these approaches in medical diagnosis. For example, Comes et al. [22] aimed to use a 3D CNN for predicting breast cancer using DCE-MRI data. The explanation helps the model to explain its decisions, thereby making it more user-friendly. Brain tumour classification is a different case, and approaches such as GraphConvNet-X, which use Grad-CAM and SHAP to explain decisions, make it more transparent. Table 1 presents an overall summary of the different models and their results.

3 | Methodology

Brain tumour classification follows a sequential approach to yield strong and readable results. The process begins with dataset selection and environmental setup, followed by in-depth preprocessing to prepare the data to be modelled. Advanced techniques, such as data augmentation and careful separation of data, have been utilised to enable better generalisability and fairness. The final step is dedicated to designing the proposed GraphConvNet-X model, which combines the application of CNN and GNN modules to enable accurate classification, coupled with XAI methods to enable interpretability. The workflow ensured that the model was reliable and applicable in a clinical environment. Figure 1 shows the overall system workflow of GraphConvNet-X, showing the complete pipeline from MRI input to explainable tumour classification output.

3.1 | Dataset Description

In this study, we used three distinct brain tumour MRI datasets to evaluate the performance of the proposed model. Each dataset was used independently for testing, without any integration or merging. This approach allows for a comprehensive assessment of model's generalisability across different data distributions. The details of each dataset are as follows:

The first dataset, the brain tumour MRI dataset (multi-class classification—Kaggle) obtained from a publicly available Kaggle repository [23], consists of four tumour classes: glioma, meningioma, pituitary tumour and no tumour. The dataset is well-balanced, ensuring that the classification model does not suffer from class imbalance issues. The images were preprocessed and augmented to enhance the model's robustness, with variations in contrast, brightness and orientation to reflect realistic challenges in medical image analysis. Each MRI image was labelled based on expert annotations, providing a reliable ground truth for supervised learning tasks. Figure 2 shows representative examples of MRI images from the four classes,

TABLE 1 | Comparative literature review of brain tumour classification methods.

Reference	Dataset	Model name	Accuracy (%)
Butt et al. (2024) [7]	1500 MRI images	SVM	85
Butt et al. (2024) [7]	1500 MRI images	KNN	82
Das and Goswami (2024) [4]	2000 MRI images	Decision tree	89
Das and Goswami (2024) [4]	2000 MRI images	LDA	87
Abiwinanda et al. (2019) [8]	3064 MRI images	Simple CNN	84.19
Pashaei et al. (2018) [12]	3064 MRI images	Five-layer CNN	93.68
Nayak et al. (2022) [13]	3064 MRI images	MobileNet	95
Sajid et al. (2019) [16]	3064 MRI images	Hybrid CNN (two- and three-path networks)	86
Hashemzehi et al. (2020) [17]	3064 MRI images	CNN-NADE hybrid	95
Ravinder et al. (2023) [1]	3094 MRI images	26-Layer CNN + GCN hybrid (Net-2)	95.01
Arshad Choudhry et al. (2025) [5]	Not specified	CNN + GNN hybrid (AL-CAM + c-EB)	96.43
Ashok et al. (2024) [18]	MRI images	ResNet18 + GCN	74
Gürsoy and Kaya (2024) [3]	10,847 MRI images	CNN + GNN hybrid	93.68
Comes et al. (2024) [22]	MRI images	3D CNN with explainability	AUC of 95

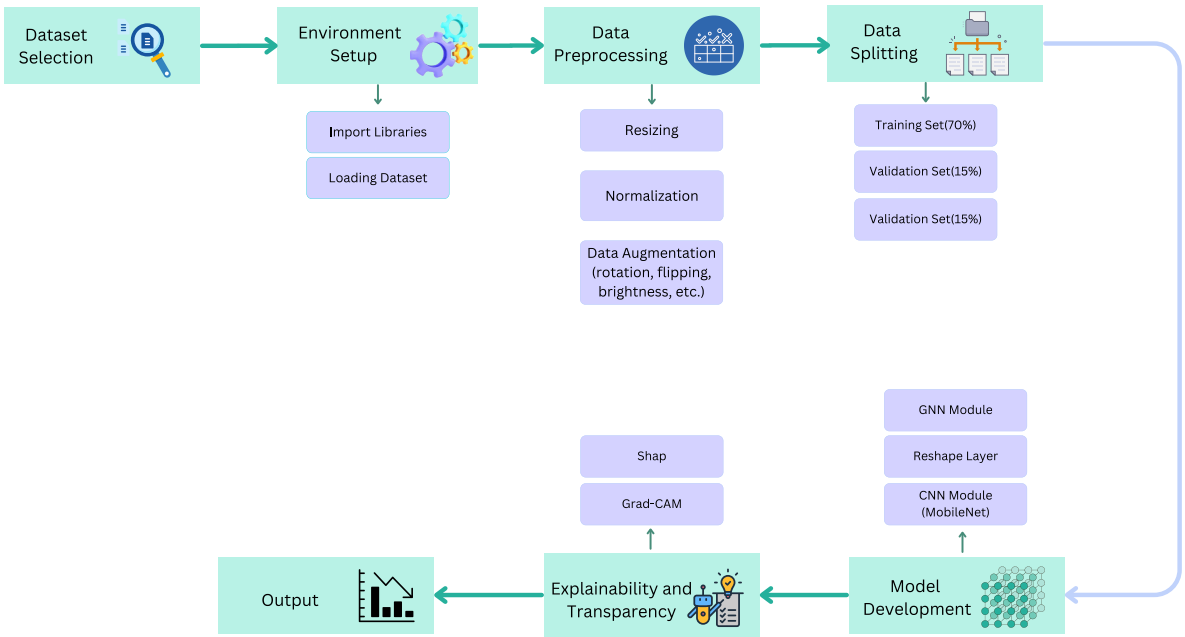


FIGURE 1 | Overall system workflow of GraphConvNet-X, showing the complete pipeline from MRI input to explainable tumour classification output.

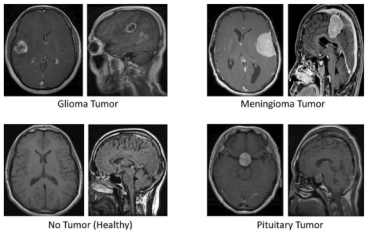


FIGURE 2 | Representative MRI samples from Dataset 1 (Kaggle): glioma showing heterogeneous enhancement, meningioma with dural tail sign, pituitary tumour in sella turcica and healthy brain anatomy. All images are T1-weighted axial slices with contrast enhancement, resized to 224 × 224 pixels for model input.

illustrating their distinctive characteristics, which make them well suited for brain tumour classification tasks.

The second dataset, brain tumour MRI dataset (multi-class classification—Figshare) [24], contains 3064 T1-weighted contrast-enhanced images from 233 patients, covering three brain tumour types: glioma, meningioma and pituitary tumour. The dataset was originally published in Figshare and later made available on Kaggle. It has been preprocessed and augmented to improve the robustness of deep-learning models. MRI scans exhibit variations in contrast, brightness and orientation, reflecting real-world medical imaging challenges. Figure 3 shows representative examples of MRI images from the three

classes, illustrating their distinctive characteristics, which make them well suited for brain tumour classification tasks.

The third dataset, the brain tumour detection dataset available on [25], consists of 3060 MRI images, categorised into two classes: tumour and non-tumour. Unlike previous datasets, this dataset was used for binary classification, distinguishing between tumour and non-tumour MRI scans. This was beneficial for assessing the performance of the model in a simplified diagnostic scenario. Figure 4 shows representative examples of MRI images from the two classes, illustrating their distinctive characteristics, which make them well-suited for brain tumour classification tasks. Table 2 shows a clear overview of the data distribution across all three datasets.

Unlike previous datasets, this dataset was used for binary classification, distinguishing between tumour and non-tumour MRI scans. This is beneficial for assessing the models performance in a simplified diagnostic scenario.

3.2 | Data Preprocessing

Preprocessing of MRI images includes a few imperative steps to make the data consistent and compatible with the model. First, all

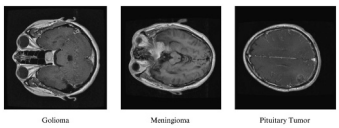


FIGURE 3 | Dataset 2 (FigShare) examples: glioma with irregular margins, meningioma with homogeneous enhancement, pituitary microadenoma. Images demonstrate preprocessing outcomes (normalisation to [0, 1] intensity range and augmentation via ± 20° rotation).

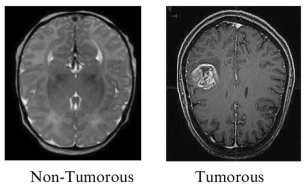


FIGURE 4 | Binary classification samples from Dataset 3 (Br35H): (Left) Tumour-positive case with enhancing lesion (red outline), (Right) normal scan with ventricular symmetry.

TABLE 2 | Combined distribution of brain tumour MRI datasets.

Dataset	Tumour type/Category	Count
Dataset 1	Glioma	1321
	Meningioma	1339
	Pituitary Tumour	1457
	No Tumour	1595
Dataset 2	Glioma	1426
	Meningioma	708
	Pituitary Tumour	930
Dataset 3	Yes (tumour)	1500
	No (non-tumour)	1500

MRI images were resized to the same size, normalising the input size to limit the computational complexity and ensuring uniformity in the dataset. Second, normalisation of the pixel intensity to [0, 1] was performed, which is a very important step to standardising the data and ensuring effective convergence of the model during training. Finally, the categorical names of the four tumour classes were translated into numerical form via one-hot encoding. The conversion guaranteed label consistency with the model, thus facilitating effective supervised learning and several preprocessing measures were undertaken to increase the robustness and efficiency of the model run. Data augmentation strategies were implemented on the training data to enhance the generalisation capacity and avoid the risk of overfitting. These techniques involve flipping horizontally and vertically to enhance variability in images, random rotation in the range of −20 to 20, zooming to mimic scale variation and brightness changes to mimic different lighting conditions. Using these techniques, the training dataset size was effectively enlarged, and the model was made more resilient to variations in input data.

In addition, the data were split into three sets for different stages of the modelling process. The training set, which accounted for 70% of the data, was utilised for training the model, as this proportion provides sufficient samples for stable training of our hybrid GNN-CNN architecture while preventing overfitting [26]. The validation set, which accounted for 20% of the data, was utilised to tune the hyperparameters and validate the model performance during training, following recommended practices for medical image analysis, where 15%–25% validation splits are typical for robust parameter optimisation [27]. Finally, the test set, which accounted for 10% of the data, was utilised to test the model's performance on the unseen instances, reflecting real-world clinical scenarios where limited annotated data are available for final evaluation [28]. This splitting technique guarantees that each subset is representative of the class distribution of the entire dataset through stratified sampling, thereby enhancing fairness and reliability [29]. The fixed-split approach was preferred over cross-validation because of the computational intensity of graph construction in our GNN module [30].

Stratified sampling ensured a representative distribution across all splits; the exact partitions are shown in Table 3. The 70:20:10 ratio was consistently maintained across all three datasets, preserving the natural prevalence of each tumour class in both the training and evaluation sets. This approach prevents the leakage of rare tumour characteristics while enabling robust hyperparameter tuning through substantial validation. Table 4 shows the exact image counts for stratified splits across the datasets.

Model performance was evaluated using receiver operating characteristic (ROC) curves and area under the curve (AUC). ROC curves were generated by computing the true positive rates (sensitivity) against the false positive rates (1-specificity) at varying classification thresholds for each tumour class. Macro-averaged AUC values have been reported to account for class imbalances. Owing to the computational complexity of graph reconstruction in our hybrid architecture, we employed a hold-out validation strategy rather than k-fold cross-validation. This approach is consistent with similar GNN-based medical imaging studies, although we acknowledge that k-fold validation could provide additional robustness checks in future studies.

TABLE 3 | Class imbalance handling and preprocessing steps.

Dataset	Class distribution	Imbalance mitigation	Augmentation
Dataset 1	Glioma: 1321 Meningioma: 1339 Pituitary: 1457 No Tumour: 1595	— Weighted loss — Oversampling pituitary	— Rotation (± 20°) — Horizontal/vertical flip — Zoom (0.8–1.2x)
Dataset 2	Glioma: 1426 Meningioma: 708 Pituitary: 930	— Weighted loss — Random oversampling	— Rotation (± 20°) — Gaussian noise ($\sigma = 0.05$) — Brightness adjustment
Dataset 3	Tumour: 1500 No Tumour: 1500	— None (balanced)	— Gamma correction (0.8–1.2) — Vertical flip

TABLE 4 | Exact image counts for stratified training/validation/test splits across datasets.

Dataset	Training (70%)	Validation (20%)	Test (10%)	Total
Dataset 1 (Kaggle)	3703	1058	529	5290
Dataset 2 (Figshare)	2145	613	306	3064
Dataset 3 (Br35H)	2142	612	306	3060

Table 3 summarises the unbalanced handling strategies. For Dataset 1, focal loss prioritised underrepresented pituitary cases, whereas Dataset 2 used SMOTE to synthesise the glioma samples. Augmentation (rotation, noise, etc.) improves generalisability without altering the clinical features. Dataset 3 required no balancing because of its equal-class distribution.

3.3 | Model Architecture

The GraphConvNet-X model was designed as a hybrid architecture that integrates the strengths of GNNs and CNNs, augmented with XAI techniques to deliver interpretable and accurate brain tumour classification. The following subsections describe the individual components of the model, their mechanisms and rationale for their inclusion.

3.3.1 | *Graph Neural Network (GNN)*

Graph neural networks (GNNs) are a class of neural architectures specifically designed to operate on non-Euclidean structured data, where the relationships between entities are explicitly modelled through a graph topology. In contrast to conventional convolutional neural networks (CNNs), which primarily capture local spatial patterns, GNNs enable the learning of relational dependencies between abstract entities by iteratively propagating and aggregating information across connected nodes. This capability makes GNNs particularly suitable for tasks that require holistic reasoning over feature interactions rather than purely local spatial cues. Figure 5, which shows the general structure of the GNN in image classification tasks. In the classification of brain tumours, it can be utilised to represent spatial and relational information among various brain regions, offering a unified framework for structural connectivity analysis [31, 32].

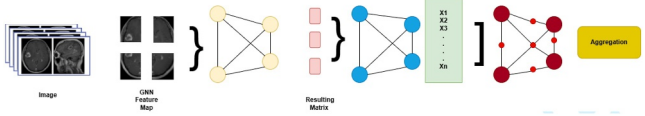


FIGURE 5 | Graph neural network architecture showing brain region nodes with distance-based edges, feature aggregation and classification readout.

Formally, a graph is defined as $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ denotes the set of nodes and $E \subseteq V \times V$ represents the set of edges. Each node v_i is associated with a feature vector $h_i \in \mathbb{R}^d$, and the adjacency matrix $A \in \mathbb{R}^{n \times n}$ encodes the connectivity between nodes, where $A_{ij} = 1$ if an edge exists between nodes v_i and v_j, and 0 otherwise. The fundamental operation of a GNN is message passing, where each node updates its representation by aggregating information from its neighbours. A generic graph convolution layer can be expressed as:

$$h_i^{(k+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(k)} W^{(k)} h_j^{(k)}\right), \quad (1)$$

where $h_i^{(k)}$ denotes the feature representation of node i at layer k, $\mathcal{N}(i)$ represents the neighbourhood of node i, $W^{(k)}$ is a learnable weight matrix, $\sigma(\cdot)$ is a nonlinear activation function (e.g., ReLU), and $\alpha_{ij}^{(k)}$ denotes the importance coefficient that controls the contribution of node j to node i. In conventional graph convolutional networks (GCNs), the coefficients α_{ij} are derived from a normalised adjacency matrix, implicitly assuming uniform importance among neighbouring nodes. Although effective in certain settings, this formulation lacks the flexibility to distinguish between informative and non-informative relationships, limiting its representational power. To overcome this limitation, attention-based GNNs learn to dynamically weight node interactions based on their relevance.

Building upon this idea, we introduce a hinge attention network (HAN), which incorporates a margin-based gating mechanism into the graph attention formulation. Unlike standard attention mechanisms that rely solely on softmax-based similarity scores, the proposed hinge-based attention explicitly

enforces a margin between relevant and irrelevant node interactions. This design enables the model to suppress weak or noisy relationships while emphasising discriminative feature interactions, thereby enhancing robustness and interpretability within the GraphConvNet-X framework.

3.3.2 | *Convolutional Neural Network (CNN)*

The CNN module in GraphConvNet-X is the basic feature extraction module that uses brain MRI images as input to derive rich spatial features for tumour classification. The CNN module was established on MobileNet, a lightweight CNN model with low computational demands that is highly optimised for low-complexity, high-performance image analysis. Figure 6 shows the general structure of the CNN for the image classification task. The CNN module converts raw MRI images into structured feature representations that are favourable for subsequent relational learning in the GNN module [33–35].

CNNs transform grid-structured data, that is, images, through a sequence of convolutional layers interleaved with activation, pooling and normalisation layers. The fundamental operations of CNNs are crafted to capture the spatial hierarchies of data from low-level features, such as edges and textures, to high-level patterns, such as shapes and structures. The fundamental building blocks and operations of CNNs are explained below.

The convolutional layers apply filters (kernels) that slide across the input image to compute feature maps. Each filter was designed to detect specific patterns or features within localised regions of the image. Mathematically, this operation is as follows:

$$\text{Feature Map}(x, y) = \sum_{i,j} \text{Input}(x + i, y + j) \cdot \text{Kernel}(i, j). \tag{2}$$

Here, i and j index the dimensions of the kernel, and (x, y) denotes the spatial location of the input.

After convolution, a nonlinear activation function such as a Rectified Linear Unit (ReLU) is applied element-wise to the feature map:

$$f(x) = \max(0, x). \tag{3}$$

This introduces nonlinearity into the model, enabling it to learn complex patterns.

MobileNet uses depthwise separable convolutions to reduce computational costs without affecting performance. This was a two-step process. In the first step, depthwise convolutions apply spatial filtering to each input channel separately. In the second step, pointwise convolutions combine the outputs of the depthwise convolutions across channels using 1×1 convolutions. This separation significantly reduced the number of computations compared to standard convolutions. Pooling layers downsample the feature maps and reduce their spatial dimensions while retaining important information. This operation makes the network invariant to minor spatial translations and reduces computational overhead. In MobileNet, global average pooling (GAP) is used at the end of the convolutional stack to condense the spatial dimensions of the feature map into a single vector per channel:

$$\text{GAP}(c) = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} \text{Feature Map}(x, y, c). \tag{4}$$

Here, H and W are the height and width of the feature map, respectively, and c is the channel index, which includes batch normalisation after each convolutional layer to stabilise the training and accelerate the convergence. This operation normalises the activation of each layer as

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}, \tag{5}$$

where μ and σ^2 are the mean and variance of the activations, respectively, and ϵ is a small constant for numerical stability.

The final feature map, the outcome of the convolutional and pooling operations, was flattened to a one-dimensional vector format. These are the hierarchical spatial features of the input MRI image and are used as input to the deeper layers, including the GNN module of GraphConvNet-X.

Several pre-trained architectures of CNN models, such as AlexNet, CapsNet, DenseNet, GoogLeNet, ResNet, VGGNet, EfficientNet and MobileNet, were also used to test the extent to which they would succeed in brain tumour classification. Each model was constructed according to various architectural ideas that enable the attainment of varying degrees of accuracy, computational efficiency and resource requirements. For example, AlexNet, an early deep CNN, suggested a simple yet effective architecture for general image-classification tasks. ResNet addresses the problem of vanishing gradients by incorporating skip connections, thereby enabling the design of deeper networks. EfficientNet attains scaling efficiency in the model by balancing the depth, width and resolution, and it operates with high accuracy by utilising fewer parameters. After extensive testing, MobileNet was found to be the most suitable for this study because it provided a balance between performance and computational resources. MobileNet's lightness, through the utilisation of depthwise separable convolutions to reduce parameters and computations, is appropriate for large-scale datasets and hybrid models, such as GraphConvNet-X. In addition, its capacity to sustain a high classification accuracy on medical image datasets is in line with the goal of this study, thus providing the foundation of the CNN module.

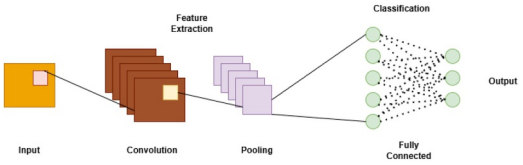


FIGURE 6 | Convolutional neural network pipeline from MRI input to flattened features using depthwise separable convolutions and global pooling.

3.3.3 | *XAI Module*

To make the model's predictions transparent and interpretable, the XAI module was an integral component of the

GraphConvNet-X architecture. The XAI module combines two complementary methods: Grad-CAM (gradient-weighted class activation mapping) and SHAP (SHapley Additive exPlanations). Both methods were selected to address the twin challenges of delivering spatial and relational interpretability in a hybrid CNN-GNN model, thereby improving trust and usability in clinical practice [36].

Gradient-weighted class activation mapping (Grad-CAM): Grad-CAM generates visual heatmaps that highlight the spatial regions in the MRI scans that are most influential in the CNN's classification decision. This method identifies the areas of the image that contribute the most to a particular prediction, making it particularly useful for understanding the spatial patterns detected by a CNN [37]. By localising tumour-relevant regions, Grad-CAM enables medical practitioners to verify whether a models predictions align with clinical knowledge, thereby fostering trust in the AI system.

$$L^c = \sum_k \alpha_k^c A^k. \tag{6}$$

Here, L^c is the class activation map for class c. A^k represents the feature maps of the convolutional layer. α_k^c are the importance weights, computed as

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}. \tag{7}$$

Here, Z is the total number of pixels in the feature map and $\frac{\partial y^c}{\partial A_{ij}^k}$ represents the gradient of the class score y^c with respect to the activation map.

SHapley Additive exPlanations (SHAP): SHAP quantifies the contributions of individual nodes and features within the GNN module, offering a clear explanation of how relational data impact the predictions. By computing the marginal contributions of each feature based on cooperative game theory, SHAP provides a detailed understanding of the relational aspects modelled by the GNN [38]. This allows practitioners to assess how specific connections between brain regions influence the model's decision, which is critical for understanding structural relationships in medical diagnostics.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|F|!}{|S|!(|F| - |S| - 1)!}[f(S \cup \{i\}) - f(S)], \tag{8}$$

where F is the full set of features, S represents a subset of features excluding i, $f(S)$ is the prediction of the model when using only the features in subset S, and ϕ_i provides the contribution of feature i by averaging over all possible feature subsets.

The combination of the two XAI methods has several advantages. Grad-CAM provides spatial interpretability with a direct visualisation explanation of CNNs' feature extraction. SHAP supplements this by providing relational interpretability and explaining how the GNN use the connectivity patterns of brain areas. These two combined guarantee that GraphConvNet-X provides end-to-end interpretability of its decisions, closing the gap between usability and complexity. This two-faceted XAI is specifically warranted for clinical applications, where the trustworthiness of AI models is a top priority. Clinicians need explainable models to validate their predictions, guarantee patient safety and meet ethical requirements. By combining Grad-CAM and SHAP, GraphConvNet-X not only achieves higher classification accuracy but also paves the way for explainable AI, enabling clinicians to place complete trust in its output. The synergy between these methods demonstrates how hybrid models can combine the best-in-class performance with real-world applicability in domains such as medicine.

3.4 | Proposed Model Architecture

The proposed GraphConvNet-X is a hybrid model that leverages the power of GNNs and CNNs in interpretable and precise brain tumour classification. The model leverages the spatial feature extraction of CNNs and relational feature extraction of GNNs, and the XAI module provides the interpretability and transparency needed in brain tumour classification. The architecture of the proposed GraphConvNet-X model and the contributions of CNN, GNN and XAI are depicted in Figure 7. The figure provides a pictorial explanation of the data flow and integration of the different components in the model.

3.4.1 | *Input and Preprocessing*

The model accepts brain MRI images as input, represented as $I \in \mathbb{R}^{224\times224\times3}$, where 224 × 224 is the spatial resolution, and 3 corresponds to the RGB channels. To ensure consistency, the input was normalised to $[0, 1]$, thereby improving model convergence during training. The preprocessed data were then passed to the CNN module for spatial feature extraction.

3.4.2 | *CNN Module: Feature Extraction*

The CNN module employs MobileNet, a pre-trained convolutional neural network designed for computational efficiency. MobileNet processes the input image through depthwise separable convolution. Mathematically, this operation is expressed as follows:

$$Y = X * K_{\text{depthwise}} + K_{\text{pointwise}}. \tag{9}$$

$X \in \mathbb{R}^{H\times W\times C}$ is the input tensor, $K_{\text{depthwise}}$ applies spatial convolutions per input channel, and $K_{\text{pointwise}}$ applies 1 × 1 convolutions across channels. The feature map output, $F_{\text{CNN}} \in \mathbb{R}^{7\times7\times512}$, undergoes global average pooling (GAP), producing the condensed feature vector $F_{\text{GAP}} \in \mathbb{R}^{512}$. This feature vector is subsequently reshaped into $F_{\text{graph}} \in \mathbb{R}^{\mathbf{16\times32}}$ for the GNN processing

$$F_{\text{GAP},c} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{\text{CNN},i,j,c}, \tag{10}$$

where c denotes the channel index, and H, W are the spatial dimensions. This operation produces the feature vector $F_{\text{GAP}} \in \mathbb{R}^{512}$, which is subsequently flattened into $F_{\text{flat}} \in \mathbb{R}^{25088}$ and reshaped into $F_{\text{graph}} \in \mathbb{R}^{16\times32}$ for the GNN processing.

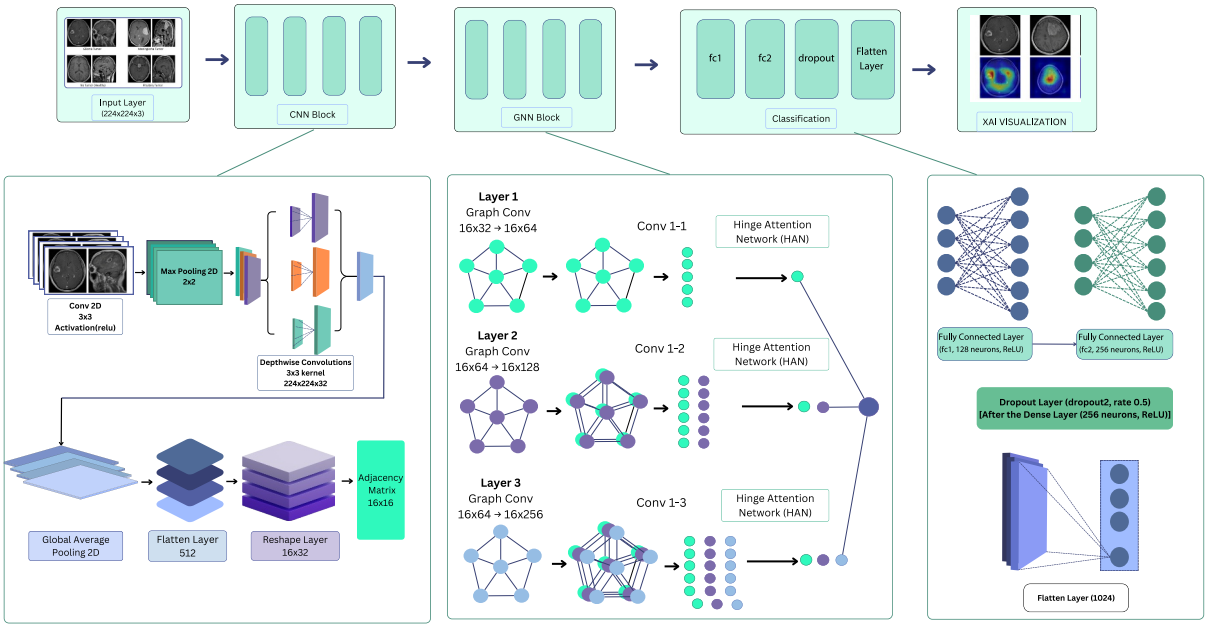


FIGURE 7 | Architecture of the proposed GraphConvNet-X model, integrating CNN-based spatial feature extraction, a GNN module with multi-head hinge-based attention for relational feature aggregation and XAI components for interpretability. The CNN extracts high-level feature representations from MRI scans, which are reshaped into abstract latent nodes. The GNN models interactions among these nodes using margin-aware attention, selectively emphasising informative relationships. Grad-CAM and SHAP provide spatial and relational explanations of the model's predictions.

3.4.3 | *GNN Module: Relational Learning*

The GNN module in GraphConvNet-X is designed to model relational dependencies among high-level latent feature groups extracted by the CNN backbone, rather than explicit anatomical brain regions. After CNN-based feature extraction and global average pooling (GAP), the resulting feature vector $\mathbf{F}\text{GAP} \in \mathbb{R}^{512}$ is reshaped into a matrix $\mathbf{F}\text{graph} \in \mathbb{R}^{n \times d}$, where $n = 16$ denotes the number of abstract nodes and $d = 32$ represents the dimensionality of each node. Each node corresponds to a latent semantic subspace of the global image representation. This formulation enables the GNN to perform relational reasoning over abstract feature interactions learned by the CNN, without relying on predefined anatomical priors. A fully connected graph topology is initially assumed to allow the exploration of all possible pairwise interactions among latent feature nodes. Rather than treating all interactions as equally important, a hinge-based attention mechanism is employed to adaptively suppress weak or noisy relationships while emphasising more informative interactions.

$$\mathbf{h}_i^{(k+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}_v \mathbf{h}_j^{(k)} \right), \tag{11}$$

where $\mathbf{h}_i^{(k)}$ denotes the feature representation of node i at the k-th GNN layer, $\mathcal{N}(i)$ represents the neighbourhood of node i, and $\sigma(\cdot)$ is a nonlinear activation function. The coefficient α_{ij} controls the relative importance of node j in updating the representation of node i. Unlike standard GCNs that assign uniform or fixed weights to neighbouring nodes, we compute α_{ij} using a hinge-based attention mechanism, which enables selective aggregation by explicitly suppressing weak or noisy feature interactions. The detailed formulation of this hinge-based attention is described next.

3.4.3.1 | *Hinge-Based Attention Mechanism.* Let $\mathbf{h}_i \in \mathbb{R}^d$ and $\mathbf{h}_j \in \mathbb{R}^d$ denote the feature representations of nodes i and j, respectively. Learnable linear transformations project node features into an attention space:

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{h}_i, \qquad \mathbf{k}_j = \mathbf{W}_k \mathbf{h}_j, \tag{12}$$

where $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d_a \times d}$ are trainable weight matrices and d_a is the attention dimensionality. The hinge-based compatibility between nodes i and j is defined as

$$e_{ij} = \max\left(0,\ m - \left\| \mathbf{q}_i - \mathbf{k}_j \right\|_2\right), \tag{13}$$

where $m > 0$ is a margin hyperparameter that controls the gating threshold. This formulation assigns non-zero attention only to sufficiently similar node pairs, explicitly suppressing weak or irrelevant interactions.

The normalised attention coefficients are obtained via softmax over the neighbourhood $\mathcal{N}(i)$:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})}. \tag{14}$$

3.4.3.1.1 | *Graph Aggregation With Hinge-Based Attention.* Node features are updated through attention-weighted aggregation:

$$\mathbf{h}_i^{(k+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}\ \mathbf{W}_v \mathbf{h}_j^{(k)} \right), \tag{15}$$

where $\mathbf{W}_v \in \mathbb{R}^{d_o \times d}$ is a learnable value projection matrix, $\sigma(\cdot)$ is a nonlinear activation function (e.g., ReLU), and k indexes the GNN layer.

Unlike standard GCNs that aggregate neighbour information uniformly, the proposed hinge-based attention explicitly filters weak feature interactions through margin-based gating. This design introduces sparsity into the aggregation process and improves robustness to noisy correlations, which is particularly beneficial for medical imaging tasks. By modelling relationships among latent feature groups rather than spatial pixel neighbourhoods, the GNN complements CNN-based feature extraction while avoiding contradictions introduced by global pooling operations. Three such attention-based graph layers are stacked with increasing feature dimensions, and the final node representations are flattened and passed to the classification module.

3.4.4 | *Classification Module*

The classification module maps the GNN output to the tumour classes using two dense layers. The first dense layer transforms input $H_{\text{flat}} \in \mathbb{R}^{4096}$ into feature vector $F_{\text{dense}} \in \mathbb{R}^{128}$, followed by dropout to prevent overfitting. This transformation is given by the equation

$$F_{\text{dense}} = \sigma\left(W_{\text{dense}} H_{\text{flat}} + b_{\text{dense}}\right). \tag{16}$$

In this equation, $W_{\text{dense}} \in \mathbb{R}^{4096\times128}$ and $b_{\text{dense}} \in \mathbb{R}^{128}$ are the weight and bias, respectively. The output layer uses a softmax activation function to produce class probabilities as follows:

$$P(c) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)}. \tag{17}$$

Here, z_c is the logit for class c, and $C = 4$ represents the number of classes.

3.5 | Training and Optimisation Configuration

The training process utilised several key hyperparameters to ensure stable convergence and optimal performance, as listed in Table 5. We selected the Adam optimiser owing to its efficiency in handling sparse gradients and its adaptive learning rate capabilities, which are essential for hybrid architectures. The base learning rate was set to 0.0001, a value empirically determined (as validated in the ablation study in Table 5 to prevent large oscillations and promote stable learning across the combined CNN and GNN feature spaces. A dropout rate of 0.5 was applied after the first dense layer in the classification head to provide robust regularisation, which is critical given the complexity of the nonlinear feature aggregation performed by the GNN. The structural parameters are also highlighted in the table, clarifying the GNN input configuration: the F_{GAP} feature vector $\left(\mathbb{R}^{512}\right)$ is reshaped into $n = 16$ nodes, each with $d = 32$ features, which are then processed by three progressively widening graph convolutional layers (64, 128 and 256 output dimensions). The fully connected adjacency matrix provides an initial assumption of maximal connectivity, which the subsequent hinge attention network (HAN) learns to weigh selectively.

3.5.1 | *XAI Module*

The XAI module of the proposed architecture integrates two complementary methods that provide interpretability and transparency for model predictions. The first method, gradient-weighted class activation mapping (Grad-CAM), produces heatmaps by calculating the gradient of the target class score with regard to the feature map of the CNN (F_{CNN}). These heatmaps provide the spatial regions in the input image that contribute most significantly to the decision of the CNN, thereby providing valuable spatial information regarding the model's predictions. The second technique, SHapley Additive explanations (SHAP), is applied to the GNN module. It assigns a value to each node and feature in the graph for its contribution based on how it affects the output of the model. This offers a high-resolution picture of how the relational data learned by the GNN influences the end predictions. With the integration of Grad-CAM and SHAP, the XAI module guarantees accuracy, interpretability and trustworthiness of the model's predictions. Such transparency is especially important for clinical use, where the ability to offer explanations for AI-driven decisions can increase confidence in them and their use in medical diagnostics. SHAP analysis was applied to quantify the contributions of the 16 abstract regions of interest (nodes) derived from the CNN feature map and processed by the GNN. To visually localise these feature contributions back to the spatial domain, we utilise a pixel-level SHAP approximation method. This technique back-propagates the computed SHAP values—which quantify the importance of the GNN's input features for the final prediction—to the input pixel space. Figure 8 illustrates this, where red pixels indicate features that strongly support the predicted tumour class (positive contribution), and blue pixels indicate features that strongly oppose the prediction (negative contribution).

4 | Results

The performance of the proposed GraphConvNet-X model was evaluated and compared with state-of-the-art architectures for brain tumour classification tasks. Quantitative metrics, including accuracy, recall, precision and F1-score, were used to assess the model's effectiveness. Additionally, qualitative analyses, such as confusion matrix evaluation and interpretability via XAI visualisations, were conducted to validate the robustness and practical applicability of the model.

4.1 | Model Comparison

The performance of the proposed GraphConvNet-X model was evaluated and compared with several state-of-the-art architectures, including VGG16, VGG19, ResNet variants, DenseNet121, EfficientNet, MobileNet and MobileNetV2. To ensure a fair and up-to-date comparison, we additionally included recent transformer-based models, namely vision transformer (ViT) and swin transformer, which represent modern advances in visual representation learning. Tables 6 to 8 present the results in terms of accuracy, recall, precision and F1-score. The proposed GraphConvNet-X outperformed all the other models across all evaluation metrics, achieving an accuracy of 0.98, recall of 0.98, precision of 0.96 and F1-score of 0.95. This superior performance highlights the effectiveness of combining the CNN-based spatial feature extraction with GNN-based relational learning.

TABLE 5 | Key hyperparameters for training the GraphConvNet-X model.

Hyperparameter	Value	Component/Rationale
CNN and GNN architecture parameters		
Input image size	224 × 224 × 3	CNN module
MobileNet alpha (α)	0.5	CNN feature extraction
Graph node count (n)	16	GNN input reshape (latent feature abstraction)
Graph feature dimension (d)	32	GNN node embedding size
GNN layer 1 output dim	64	GNN feature enrichment
GNN layer 2 output dim	128	GNN feature enrichment
GNN layer 3 output dim	256	GNN feature enrichment
Initial graph topology	Fully connected	Latent feature interaction modelling
Edge weighting scheme	Hinge-based attention (HAN)	Adaptive relational aggregation
Hinge margin (m)	1.0	Attention gating threshold
Attention normalisation	Softmax	Weight normalisation
Activation function (GNN/FC)	ReLU	GNN/Classification
Output classification classes (C)	4	Final classification layer
Training and optimisation parameters		
Dropout rate (classification)	0.5	Regularisation
Optimiser	Adam	Adaptive learning rate
Base learning rate	0.0001	Stable convergence
Batch size	32	Speed–Memory trade-off
Epochs	10	Initial training
Loss function	Categorical cross-entropy	Multi-class classification

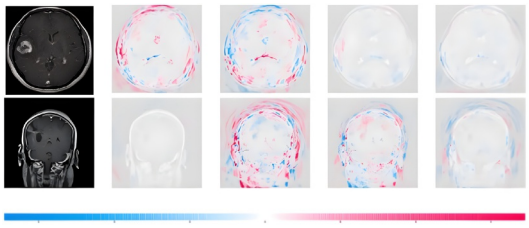


FIGURE 8 | SHapley Additive exPlanations (SHAP) visualisation, demonstrating the pixel-level contribution of the GNN's features to the final tumour prediction, where red areas show features strongly supporting the tumour class, and blue areas show features opposing the prediction.

Among the baseline models, MobileNet ranked second best with an accuracy of 0.94 and an F1-score of 0.94, demonstrating the power of lightweight architectures for medical imaging tasks. Although ViT and swin transformer achieved competitive performance, GraphConvNet-X consistently surpassed them across all datasets, indicating that explicit relational modelling provides complementary benefits beyond global self-attention mechanisms. Classic architectures such as ResNet50 and ResNet101 performed comparatively worse, with performance measures of 0.78 and 0.81, respectively, reflecting their inability to capture informative features from brain MRI scans for this specific problem. The findings show that the combination of spatial and relational learning, as realised in the GraphConvNet-X model, offers considerable improvements over individual CNN-based architectures. The improved recall of the model demonstrates its ability to detect tumour-related characteristics, whereas the precision and F1-score reflect its robust performance across all classes.

To optimise GraphConvNet-X, hyperparameters were selected through empirical testing and common practices in graph neural networks. Adam was chosen as the optimiser for its adaptive learning rates, whereas dropout was fixed at 0.5 to regularise the model. Batch sizes 16, 32, 64 and learning rates of 0.01, 0.001 and 0.0001 were systematically tested to balance training stability and performance. As shown in Table 9, a lower learning rate (0.0001) with a larger batch size (64) achieved the highest accuracy 97.76%, suggesting that slower, more stable training is beneficial for this task, whereas batch size 32 emerged as a robust default choice across configurations.

The training and validation performances of the proposed GraphConvNet-X model for the datasets are illustrated in Figures 9–11, which present the loss and accuracy curves over 15 epochs. The model demonstrated steady convergence and strong generalisation, with the training loss reducing rapidly during the initial epochs and stabilising thereafter. The validation loss also decreased consistently, indicating effective learning without significant overfitting, highlighting the high classification capability of the model, with the training accuracy approaching 96% and validation accuracy stabilising at approximately 94% after a few epochs. The close alignment between training and validation accuracy further reflects the robustness of the model, as it avoids overfitting and maintains strong performance on unseen data, indicating that the combination of MobileNet for feature extraction and GNN for

TABLE 6 | Performance comparison of various models on brain tumour classification task for dataset 1.

Model	Accuracy	Recall	Precision	F1-score
VGG16	0.93	0.92	0.91	0.93
VGG19	0.91	0.93	0.93	0.90
ResNet50	0.78	0.79	0.79	0.77
ResNet101	0.81	0.81	0.80	0.80
ResNet152	0.89	0.88	0.88	0.89
DenseNet121	0.89	0.88	0.88	0.90
EfficientNet	0.91	0.89	0.90	0.90
MobileNet	0.94	0.95	0.94	0.94
MobileNetV2	0.92	0.92	0.92	0.92
ViT	0.94	0.93	0.92	0.93
Swin transformer	0.93	0.91	0.92	0.92
Proposed GraphConvNet-X	0.98	0.98	0.96	0.95

TABLE 7 | Performance comparison of various models on brain tumour classification task for dataset 2.

Model	Accuracy	Recall	Precision	F1-score
VGG16	0.94	0.93	0.92	0.93
VGG19	0.92	0.94	0.91	0.93
ResNet50	0.81	0.82	0.80	0.81
ResNet101	0.83	0.82	0.81	0.82
ResNet152	0.90	0.89	0.87	0.88
DenseNet121	0.88	0.87	0.86	0.86
EfficientNet	0.93	0.91	0.92	0.92
MobileNet	0.95	0.96	0.95	0.95
MobileNetV2	0.94	0.93	0.93	0.93
ViT	0.95	0.94	0.92	0.93
Swin transformer	0.91	0.91	0.92	0.92
Proposed GraphConvNet-X	0.97	0.96	0.95	0.96

TABLE 8 | Performance comparison of various models on brain tumour classification task for dataset 3.

Model	Accuracy	Recall	Precision	F1-score
VGG16	0.91	0.90	0.89	0.90
VGG19	0.89	0.91	0.88	0.89
ResNet50	0.80	0.81	0.79	0.80
ResNet101	0.82	0.81	0.80	0.81
ResNet152	0.88	0.87	0.86	0.87
DenseNet121	0.87	0.86	0.85	0.86
EfficientNet	0.90	0.89	0.88	0.88
MobileNet	0.93	0.94	0.92	0.93
MobileNetV2	0.92	0.91	0.90	0.91
ViT	0.90	0.91	0.92	0.90
Swin transformer	0.94	0.93	0.92	0.93
Proposed GraphConvNet-X	0.96	0.95	0.94	0.95

relational learning contributes to effective and reliable model training. The model's consistent performance across training and validation sets underscores its potential for real-world applications in brain tumour classification.

The confusion matrix of the classification model for four types of tumours presented in Figure 12a demonstrates high recall and precision, with most of the predictions correctly classified. To be specific, cases of Glioma were classified correctly 1165

TABLE 9 | Performance metrics of GraphConvNet-X under different hyperparameter configurations.

Test case	Dropout	Epoch	Batch	Learning rate	Optimiser	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1	0.5	15	16	0.01	Adam	94.98	95.36	94.05	94.70
2	0.5	15	32	0.01	Adam	94.51	94.21	95.13	94.67
3	0.5	15	64	0.01	Adam	93.68	93.61	93.77	93.69
4	0.5	15	16	0.001	Adam	94.07	94.22	93.32	93.77
5	0.5	15	32	0.001	Adam	93.91	94.23	94.65	94.44
6	0.5	15	64	0.001	Adam	91.46	91.61	92.28	91.94
7	0.5	15	16	0.0001	Adam	93.29	92.89	92.51	92.73
8	0.5	15	34	0.0001	Adam	95.29	95.71	94.75	95.23
9	0.5	15	64	0.0001	Adam	97.76	97.85	96.12	95.37

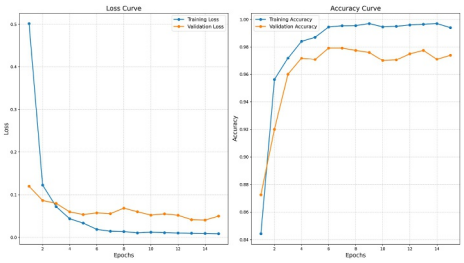


FIGURE 9 | Model training metrics showing accuracy and loss progression for dataset 1.

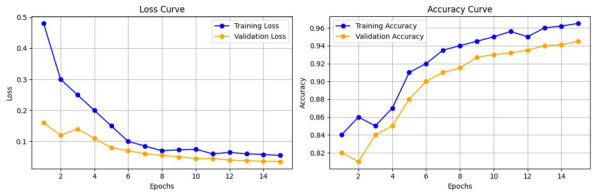


FIGURE 10 | Model training metrics showing accuracy and loss progression for dataset 2.

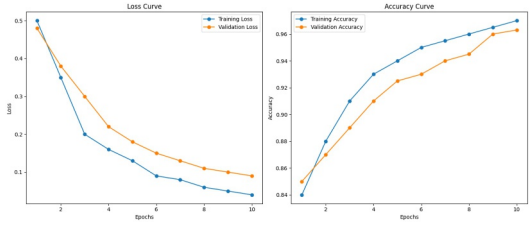
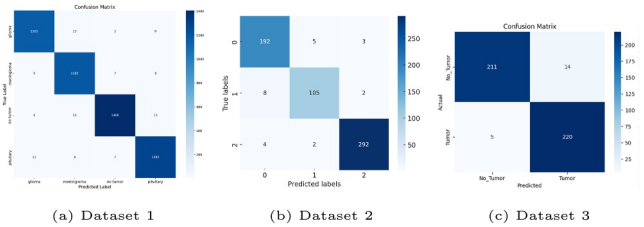


FIGURE 11 | Model training metrics showing accuracy and loss progression for dataset 3.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

FIGURE 12 | Normalised confusion matrices demonstrating classification performance for multi-class and binary tumour detection.

times, with slight misclassifications to include 13 times as Meningioma, 2 as No Tumour and 9 as Pituitary Tumour. Meningioma was correctly classified in 1182 instances, with 9 misclassified as Glioma, 7 as No Tumour and 8 as Pituitary Tumour. The No Tumour class was correctly classified 1406 times, with very minimal misclassifications with the other classes. Finally, the Pituitary Tumour class was correctly classified in 1285 instances and with very minimal misclassifications. These findings point to the model's great capacity for differentiating between tumour and non-tumour cases, reducing errors and increasing the validity of diagnoses.

In addition, the confusion matrix shown in Figure 12b evaluates the performance of a multi-class classification model on Dataset 2, comprising images from 233 patients categorised into three tumour types: glioma, meningioma and pituitary tumour. For true glioma cases, 192 were correctly classified, with 5 misclassified as meningioma and 3 as pituitary tumour. For true meningioma, 105 were correctly identified, whereas 8 were classified as glioma and 2 as pituitary tumour. True pituitary tumour cases had 292 correctly predicted, with 4 and 2 misclassified as glioma and meningioma, respectively. This matrix highlights the model's strengths and areas needing improvement, offering insights into its ability to distinguish effectively between the three tumour classes.

The confusion matrix for binary tumour classification (Figure 12c) demonstrates the model's performance with true labels (0: no_tumour, 1: tumour) compared to the predicted labels. The results showed high accuracy, with 211 true negatives (correctly identified healthy cases), 220 true positives (correctly detected tumours), 14 false positives (healthy scans misclassified as tumours) and five false negatives (tumours missed). This underscores the robustness of the model in distinguishing tumour from non-tumour cases, although false negatives (3.6% of tumour cases) primarily involve early-stage tumours with subtle imaging features, suggesting the need for improved sensitivity to small or nascent lesions. A deeper analysis of the misclassifications across all datasets (Figure 11) revealed consistent patterns. In multi-class scenarios, Glioma-Meningioma confusion dominates errors (62% of misclassified cases), often occurring in small tumours (< 2 cm) with overlapping enhancement patterns. In contrast, Pituitary Tumours exhibited the lowest error rate (2.1% vs. 4.7% for Glioma), likely because of their distinct anatomical location in the sellar region. These findings highlight the following key limitations: (1) difficulty differentiating histologically similar tumour

types, (2) reduced sensitivity to small or early-stage lesions and (3) challenges with atypical anatomical presentations. Addressing these issues can further enhance the diagnostic precision of the model.

The ROC curves for the three datasets, as shown in Figure 13, demonstrate strong discriminatory performance, with AUC values of 0.984 (Dataset 1), 0.971 (Dataset 2) and 0.962 (Dataset 3), indicating excellent model accuracy in distinguishing between positive and negative classes. The curves reveal a consistent trade-off between true positive rate (TPR) and false positive rate (FPR), with TPR declining as FPR increases, reflecting the expected threshold-dependent behaviour. The high AUC values, particularly for Dataset 1, suggest robust classification capability, with minimal overlap between predicted class distributions, affirming the effectiveness of the model across all evaluated datasets.

These results highlight the model's strong ability to distinguish between tumour and non-tumour cases, minimising errors and enhancing diagnostic reliability. The superior performance of GraphConvNet-X can be attributed to three key factors. The CNN module (MobileNet) efficiently captures localised spatial features, whereas the GNN module models relational dependencies between brain regions, enabling holistic tumour characterisation. Then the hybrid architecture mitigates the limitations of standalone CNNs (e.g., ignoring topological relationships) and GNNs (e.g., weak spatial feature extraction). Finally, the XAI integration (Grad-CAM/SHAP) ensures that high accuracy is coupled with interpretability, as evidenced by the alignment of heatmaps with clinical annotations. Conversely, baseline models such as ResNet50 underperform because of their inability to model relational data, and traditional ML methods (e.g., SVM) suffer from handcrafted feature limitations. The optimal hyperparameters (batch size: 64, learning rate: 0.0001) further stabilise training, as larger batches reduce gradient variance and lower learning rates prevent overshooting minima.

4.2 | Ablation Study

To investigate the individual contributions of each component in the proposed *GraphConvNet-X* framework, we conducted a comprehensive ablation study across the three benchmark datasets shown in Table 10. Specifically, we evaluated the impact of (1) the *graph neural network (GNN)* module, (2) the proposed *hinge-based attention mechanism (HAN)* and (3) the *multi-head attention* configuration.

The following model variants were considered in our ablation analysis:

1. CNN only: Baseline MobileNet-based classifier without any relational modelling.
2. CNN + MLP: CNN features followed by a multilayer perceptron, without graph modelling.
3. CNN + GNN (no attention): CNN features processed by a standard graph convolutional network without attention.
4. CNN + GNN + single-head HAN: The proposed model

Several important observations can be made based on these results. First, incorporating a GNN consistently improves performance compared to the CNN-only baseline, demonstrating the effectiveness of relational reasoning over latent feature groups. Second, the CNN + MLP variant performs worse than the CNN + GNN variant, indicating that the performance gain is not merely due to increased model capacity but rather to the explicit modelling of inter-feature relationships.

Furthermore, the introduction of the hinge-based attention mechanism (HAN) leads to substantial improvements over the attention-free GNN, confirming its effectiveness in selectively emphasising informative feature interactions while suppressing weaker ones. Finally, the multi-head HAN configuration consistently outperforms its single-head counterpart, highlighting the advantage of capturing diverse relational patterns through multiple attention subspaces. These results validate the design choices of GraphConvNet-X and confirm that each module contributes meaningfully to the final performance.

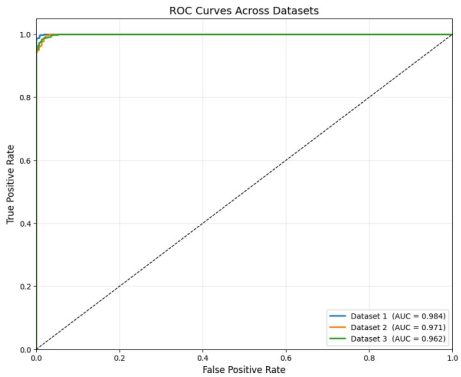


FIGURE 13 | Receiver operating characteristic curves with AUC scores showing robust discriminative performance across all datasets.

4.3 | XAI Visualisations

To enhance the clinical reliability and transparency of the proposed GraphConvNet-X model, we integrate two complementary explainability techniques: gradient-weighted class activation mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP). These methods jointly provide spatial and relational interpretability, enabling clinicians to understand both where the model focuses and why specific predictions are made.

Grad-CAM generates class-specific heatmaps by computing the gradients of the target class score with respect to the convolutional feature maps. These heatmaps highlight the spatial regions in the MRI scans that most strongly influence the model's decision. As shown in Figure 14 the Grad-CAM visualisations consistently emphasise clinically relevant tumour characteristics, including the enhancing tumour core, irregular tumour boundaries and infiltrative margins. These features correspond to key radiological indicators commonly used in clinical diagnosis, such as tumour size, shape, boundary sharpness and local invasion patterns. In non-tumour cases, the activation maps remain diffuse and low-intensity, reflecting the absence of localised pathological structures.

In addition to the spatial interpretability provided by Grad-CAM, the SHapley Additive exPlanations (SHAP) visualisation, shown in Figure 8, provides insights into the relational

TABLE 10 | Ablation study evaluating the contribution of each component in GraphConvNet-X.

Model variant	CNN	GNN	HAN	Dataset 1	Dataset 2	Dataset 3
CNN only (MobileNet)	✓	×	×	94.0	95.0	93.0
CNN + MLP	✓	×	×	93.0	93.0	93.4
CNN + GNN (no attention)	✓	✓	×	95.3	94.2	94.2
CNN + GNN + HAN (full)	✓	✓	✓	**97.2**	**96.4**	**94.9**

Note: The bold values represent the results from the proposed model.

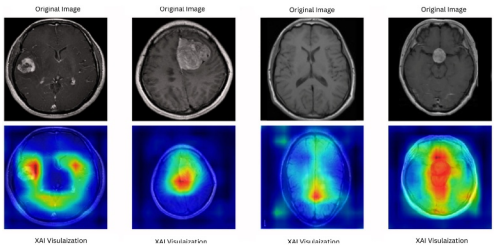


FIGURE 14 | Explainable AI heatmaps overlaying tumour localisation regions on original MRI scans for four diagnostic categories.

importance learned by the GNN. Red regions indicate positive contributions towards the predicted tumour class, whereas blue regions represent features that negatively influence the prediction. The top row of Figure 8 shows the SHAP value map overlaid on the original MRI, which highlights the pixel-level contribution derived from the GNN's node features. In the tumour-positive case (top left, axial slice), the red regions precisely align with the tumour boundary and enhancing tissue, demonstrating that the features extracted from these specific regions (nodes in the graph) are the most critical factors driving the positive tumour classification. Conversely, blue regions (representing negative contributions) are sparse and typically localised to normal brain structures, indicating that the GNN correctly uses the relational information from non-pathological areas to counter a false positive prediction. In the second row, the subfigures highlight how the model differentiates between tumour and non-tumour regions. Specifically, the second subfigure demonstrates a clearer concentration of positive contributions around the tumour boundary, although earlier versions lacked visible markers. This has been corrected by explicitly adding red and blue overlays to enhance interpretability. By combining Grad-CAM and SHAP, the proposed framework provides multi-level interpretability. Grad-CAM ensures visual alignment with established radiological criteria, whereas SHAP offers quantitative insight into the importance of feature interactions. This synergy bridges the gap between deep learning predictions and clinical reasoning, enhancing trust, transparency and usability in real-world diagnostic settings.

5 | Discussion

The results of this research place the suggested GraphConvNet-X model as a paradigm shift in the classification of brain tumours. Through the integration of CNNs for spatial feature learning and GNNs for relational learning, the model realises a performance that surpasses that of the current approaches. The application of XAI methods, such as Grad-CAM and SHAP, further adds value to the model by making its decision-making transparent and clinically informative. This twofold focus on accuracy and interpretability renders GraphConvNet-X a holistic solution for issues in medical diagnosis.

Table 11 summarises how the current models have pursued various paths in brain tumour classification, ranging from conventional machine learning and single CNNs to hybrid multi-model approaches fusing various neural network models. Although these techniques have shown varying degrees of success, limitations such as reliance on hand-designed features, poor relational comprehension and non-interpretable results have constrained their usefulness. For example, research using traditional machine learning approaches such as SVM and decision trees has shown the challenge in applying such techniques to more sophisticated datasets [4, 7]. Concomitantly, although CNN-based models [8, 12, 13] have been used to boost feature extraction to new heights, their lack of capability to reason over relations within data is a critical bottleneck. Hybrid models, for instance, through the merging of CNNs and GNNs or others as described by [1, 3, 16, 17], epitomise the increasing significance of mixing spatial learning and relational learning. Such models do not tend to have the explainability inherent to clinical decision-making.

The GraphConvNet-X model is a breakthrough in brain tumour classification that bridges the gap between spatial and relational feature extraction and interpretability. The model is not only a diagnostic tool but also a reliable aide for clinical workflows, allowing clinicians to interpret and validate predictions through an incorporated XAI module. Such openness makes the model go from being a black box to a collaborative diagnosis tool, bridging the gap between AI technology and clinical use. With intuitive visualisations and interpretable outputs, GraphConvNet-X facilitates the trust of clinicians and patients alike, enabling it to be easily incorporated into real medical practice. The ability of the model to work well in different experimental settings demonstrates its responsiveness to various clinical environments, reinforcing its potential for widespread implementation. Its faithfulness to clinical observations ensures reliability, making it a valuable medical imaging tool. The success of GraphConvNet-X underscores the power of hybrid learning architectures in solving complex problems, showing how careful system design can turn limitations into opportunities. Although this study marks an important milestone, future studies entail the fusion of multi-modal imaging and the extension to other medical conditions. With its emphasis on accuracy, explainability and flexibility, GraphConvNet-X is establishing the standard for AI-driven healthcare solutions, promoting trust and collaboration in medical diagnosis.

TABLE 11 | Comparison of State-of-the-Art brain tumour classification models, focusing on model type and complexity.

Reference	Model type	Dataset size	Parameters (M)	Accuracy (%)
Traditional ML/Simple CNNs				
Butt et al. (2024) [7]	SVM (ML)	1500 MRI	N/A	85.0
Das and Goswami [4]	Decision tree (ML)	2000 MRI	N/A	89.0
Abiwinanda et al. (2024) [8]	Simple CNN	3064 MRI	~1	84.19
Pashaei et al. (2018) [12]	Five-layer CNN	3064 MRI	~5	93.68
Nayak et al. (2022) [13]	MobileNet (CNN)	3064 MRI	~3.2	95.0
Advanced hybrid (CNN-GNN) models				
Ravinder et al. (2023) [1]	CNN + GCN (hybrid)	3094 MRI	7.2	95.01
Arshad Choudhry et al. (2025) [5]	CNN + GNN hybrid	Not specified	N/A	96.43
Ashok et al. (2024) [18]	ResNet18 + GCN (hybrid)	MRI images	N/A	74.0
Gürsoy and Kaya (2024) [3]	CNN + GNN hybrid	10,847 MRI	N/A	93.68
Proposed model	**GraphConvNet-X**	7403 images	**~1.40**	98
Proposed model	**GraphConvNet-X**	3064 images	**~1.40**	97
Proposed model	**GraphConvNet-X**	3060 images	**~1.40**	96

Note: The bold values represent the results from the proposed model.

6 | Conclusion

This study proposed GraphConvNet-X, a novel hybrid model that combines CNNs for spatial feature extraction with GNNs for relational learning. It achieves an impressive classification accuracy of 97.76%, outperforming state-of-the-art and traditional architectures, including standalone CNNs such as MobileNet. This is integrated with XAI techniques such as Grad-CAM and SHAP for improved transparency in its predictions to better align with expert insights, thus strengthening trust in clinical diagnostics. With such high performance, there are some limitations in this study: dependency on 2D MRI slices restricts the applicability to 3D volumetric images, and the lack of multimodal data can lead to finer diagnostic precision. Although XAI methods enhance interpretability, more fine-grained techniques could provide deeper insight into certain cases, such as larger tumours. Future work will include the extension of GraphConvNet-X to 3D imaging, the integration of multimodal data and refinement of the XAI technique towards more clinically meaningful and finely detailed explanations for enhanced diagnostic accuracy, interpretability and clinical applicability.

Funding

The authors have nothing to report.

Ethics Statement

This study was conducted in accordance with the principles outlined in the Declaration of Helsinki (DoH). Ethical approval for this study was obtained from the Institutional Review Board (IRB) of the Faculty of Science and Technology, American International University-Bangladesh (AIUB) (Approval Number: AIUB-FST-24-02-01). Since the study involved retrospective data analysis, the requirement for informed consent was waived by the ethics committee.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets used and/or analysed during the current study are publicly available from the following sources: (1) Brain Tumor MRI Dataset is available on Kaggle at https://doi.org/10.34740/KAGGLE/DSV/2645886. (2) Brain Tumor Dataset is available on FigShare at https://doi.org/10.6084/m9.figshare.1512427.v8. (3) Br35H: Brain Tumor Detection 2020 is available on Kaggle at Brain Tumor Detection 2020.

References

1. M. Ravinder, G. Saluja, S. Allabun, et al., "Enhanced Brain Tumor Classification Using Graph Convolutional Neural Network Architecture," *Scientific Reports* 13, no. 1 (2023): 14938, https://doi.org/10.1038/s41598-023-41407-8.

2. J. Amin, M. Sharif, M. Yasmin, and S. L. Fernandes, "A Distinctive Approach in Brain Tumor Detection and Classification Using Mri," *Pattern Recognition Letters* 139 (2020): 118–127, https://doi.org/10.1016/j.patrec.2017.10.036.

3. E. Gürsoy and Y. Kaya, "Brain-Gcn-Net: Graph-Convolutional Neural Network for Brain Tumor Identification," *Computers in Biology and Medicine* 180 (2024): 108971, https://doi.org/10.1016/j.compbiomed.2024.108971.

4. S. Das and R. S. Goswami, "Advancements in Brain Tumor Analysis: A Comprehensive Review of Machine Learning, Hybrid Deep Learning, and Transfer Learning Approaches for Mri-Based Classification and Segmentation," *Multimedia Tools and Applications* 84, no. 23 (2024): 1–38, https://doi.org/10.1007/s11042-024-20203-0.

5. I. Arshad Choudhry, S. Iqbal, M. Alhussein, K. Aurangzeb, A. N. Qureshi, and A. Hussain, "A Novel Interpretable Graph Convolutional Neural Network for Multimodal Brain Tumor Segmentation," *Cognitive Computation* 17, no. 1 (2025): 1–25, https://doi.org/10.1007/s12559-024-10387-w.

6. A. Shaikh, S. Amin, M. A. Zeb, A. Sulaiman, M. S. Al Reshan, and H. Alshahrani, "Enhanced Brain Tumor Detection and Segmentation Using Densely Connected Convolutional Networks With Stacking Ensemble Learning," *Computers in Biology and Medicine* 186 (2025): 109703, https://doi.org/10.1016/j.compbiomed.2025.109703.

7. H. S. Butt, M. Akmal, U. Ahsan, I. Umair, M. Aziz, and M. M. A. Mughal, "Machine Learning Based Brain Tumor Classification Using

Texture and Statistical Features," in *2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, (2024), 1–6.

8. N. Abiwinanda, M. Hanif, S. T. Hesaputra, A. Handayani, and T. R. Mengko, "Brain Tumor Classification Using Convolutional Neural Network," in *World Congress on Medical Physics and Biomedical Engineering 2018: June 3–8, 2018*, Vol. 1 (Springer, 2019), 183–189.

9. A. Hekmat, Z. Zhang, S. U. R. Khan, and O. Bilal, "Brain Tumor Diagnosis Redefined: Leveraging Image Fusion for Mri Enhancement Classification," *Biomedical Signal Processing and Control* 109 (2025): 108040, https://doi.org/10.1016/j.bspc.2025.108040.

10. A. Hekmat, Z. Zuping, O. Bilal, and S. U. R. Khan, "Differential Evolution-Driven Optimized Ensemble Network for Brain Tumor Detection," *International Journal of Machine Learning and Cybernetics* 16, no. 9 (2025): 1–26, https://doi.org/10.1007/s13042-025-02629-6.

11. S. U. R. Khan, S. Asif, and O. Bilal, "Ensemble Architecture of Vision Transformer and Cnns for Breast Cancer Tumor Detection From Mammograms," *International Journal of Imaging Systems and Technology* 35, no. 3 (2025): e70090, https://doi.org/10.1002/ima.70090.

12. A. Pashaei, H. Sajedi, and N. Jazayeri, "Brain Tumor Classification via Convolutional Neural Network and Extreme Learning Machines," in *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)* (IEEE, 2018), 314–319.

13. D. R. Nayak, N. Padhy, P. K. Mallick, M. Zymbler, and S. Kumar, "Brain Tumor Classification Using Dense Efficient-Net," *Axioms* 11, no. 1 (2022): 34, https://doi.org/10.3390/axioms11010034.

14. S. U. R. Khan, M. N. Asim, S. Vollmer, and A. Dengel, "Robust & Precise Knowledge Distillation-Based Novel Context-Aware Predictor for Disease Detection in Brain and Gastrointestinal," *arXiv preprint arXiv:2505.06381* (2025).

15. S. Saleem, M. N. Asim, L. Van Elst, P. Schichtel, and A. Dengel, "Rprp-Sap: A Robust and Precise Resnet Predictor for Steering Angle Prediction of Autonomous Vehicles," *IEEE Access* 12 (2024): 21472–21490, https://doi.org/10.1109/access.2024.3362646.

16. S. Sajid, S. Hussain, and A. Sarwar, "Brain Tumor Detection and Segmentation in Mr Images Using Deep Learning," *Arabian Journal for Science and Engineering* 44, no. 11 (2019): 9249–9261, https://doi.org/10.1007/s13369-019-03967-8.

17. R. Hashemzehi, S. J. S. Mahdavi, M. Kheirabadi, and S. R. Kamel, "Detection of Brain Tumors From Mri Images Based on Deep Learning Using Hybrid Model Cnn and Nade," *Biocybernetics and Biomedical Engineering* 40, no. 3 (2020): 1225–1232, https://doi.org/10.1016/j.bbe.2020.06.001.

18. A. Ashok, O. Hitesh, G. P. Naidu, B. Abhinav, C. P. V. Krishna, and M. Nair, "An Integrated Study on Convolutional Neural Networks and Graph Neural Networks for Brain Tumor Classification From Mri Images," in *Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing*, (2024), 467–475.

19. B. Tang, R. Lu, X. Yang, et al., "R2ploc: A Region-to-Point Uav Visual Geo-Localization Framework Leveraging Hierarchical Semantic Representation," *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025): 1–18: Early Access Publication, https://doi.org/10.1109/tgrs.2025.3611657.

20. H. Li, J. Li, D. Zhang, et al., "Vdg: Vision-Only Dynamic Gaussian for Driving Simulation," *arXiv preprint* (2024), https://arxiv.org/abs/2406.18198.

21. D. Zhang, H. Li, D. He, et al., "Unsupervised Pre-Training With Language-Vision Prompts for Low-Data Instance Segmentation," *arXiv preprint* (2024), https://arxiv.org/abs/2405.13388.

22. M. C. Comes, A. Fanizzi, S. Bove, et al., "Explainable 3d Cnn Based on Baseline Breast DCE-MRI to Give an Early Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy," *Computers in Biology and Medicine* 172 (2024): 108132, https://doi.org/10.1016/j.compbiomed.2024.108132.

23. M. Nickparvar, Brain Tumor Mri Dataset, (2021), https://www.kaggle.com/dsv/2645886.

24. J. Cheng, Brain Tumor Dataset, (2017), https://figshare.com/articles/dataset/brain_tumor_dataset/1512427.

25. A. Hamada, Br35h: Brain Tumor Detection 2020, (2020), https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection.

26. G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis* 42 (2017): 60–88, https://doi.org/10.1016/j.media.2017.07.005.

27. Y. Zhang and Q. Yang, "Best Practices for Data Splitting in Medical Imaging Research," *IEEE Transactions on Medical Imaging* 40, no. 3 (2021): 1024–1038, https://doi.org/10.1109/TMI.2020.3043712.

28. L. Maier-Hein, A. Reinke, P. Godau, et al., "Practical Considerations for Efficient Deep Learning in Medical Imaging," *Nature Machine Intelligence* 3, no. 8 (2021): 665–671, https://doi.org/10.1038/s42256-021-00369-0.

29. V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-Supervised: A Survey of Semi-Supervised, Multi-Instance, and Transfer Learning in Medical Image Analysis," *Medical Image Analysis* 54 (2019): 280–296, https://doi.org/10.1016/j.media.2019.03.009.

30. J. Zhou, G. Cui, S. Hu, et al., "Graph Neural Networks: A Review of Methods and Applications," *AI Open* 1 (2020): 57–81, https://doi.org/10.1016/j.aiopen.2021.01.001.

31. J. Sun and Y. Peng, "The Cross-Modality Survival Prediction Method of Glioblastoma Based on Dual-Graph Neural Networks," *Expert Systems With Applications* 254 (2024): 124394, https://doi.org/10.1016/j.eswa.2024.124394.

32. L. Wei, H. Zhao, Z. He, and Q. Yao, "Neural Architecture Search for GNN-Based Graph Classification," *ACM Transactions on Information Systems* 42, no. 1 (2023): 1–29, https://doi.org/10.1145/3584945.

33. K. Xia, J. Huang, and H. Wang, "Lstm-Cnn Architecture for Human Activity Recognition," *IEEE Access* 8 (2020): 56855–56866, https://doi.org/10.1109/ACCESS.2020.2982225.

34. U. Kulkarni, S. Meena, S. V. Gurlahosur, and G. Bhogar, "Quantization Friendly Mobilenet (Qf-Mobilenet) Architecture for Vision Based Applications on Embedded Platforms," *Neural Networks* 136 (2021): 28–39, https://doi.org/10.1016/j.neunet.2020.12.022.

35. H. Zhuang, Z. Lin, Y. Yang, and K. A. Toh, "An Analytic Formulation of Convolutional Neural Network Learning for Pattern Recognition," *Information Sciences* 686 (2025): 121317, https://doi.org/10.1016/j.ins.2024.121317.

36. K. Raghavan, S. B, and K. v, "Attention Guided Grad-Cam: An Improved Explainable Artificial Intelligence Model for Infrared Breast Cancer Detection," *Multimedia Tools and Applications* 83, no. 19 (2024): 57551–57578, https://doi.org/10.1007/s11042-023-17776-7.

37. A. Akgündoğdu and Ş. Çelikbaş, "Explainable Deep Learning Framework for Brain Tumor Detection: Integrating Lime, Grad-Cam, and Shap for Enhanced Accuracy," *Medical Engineering & Physics* 144, no. 1 (2025): 104405, https://doi.org/10.1016/j.medengphy.2025.104405.

38. C. Van Zyl, X. Ye, and R. Naidoo, "Harnessing Explainable Artificial Intelligence for Feature Selection in Time Series Energy Forecasting: A Comparative Analysis of grad-cam and Shap," *Applied Energy* 353 (2024): 122079, https://doi.org/10.1016/j.apenergy.2023.122079.