algorithms

MDPI

Article

GATransformer: A Graph Attention Network-Based Transformer Model to Generate Explainable Attentions for Brain Tumor Detection

Sara Tehsin, Inzamam Mashood Nasir and Robertas Damaševičius *

Centre of Real Time Computer Systems, Kaunas University of Technology, 51368 Kaunas, Lithuania
* Correspondence: robertas.damasevicius@ktu.lt

Abstract: Brain tumors profoundly affect human health owing to their intricacy and the difficulties associated with early identification and treatment. Precise diagnosis is essential for effective intervention; nevertheless, the resemblance among tumor forms often complicates the identification of brain tumor types, particularly in the early stages. The latest deep learning systems offer very high classification accuracy but lack explainability to help patients understand the prediction process. GATransformer, a graph attention network (GAT)-based Transformer, uses the attention mechanism, GAT, and Transformer to identify and preserve key neural network channels. The channel attention module extracts deeper properties from weight-channel connections to improve model representation. Integrating these elements results in a reduction in model size and enhancement in computing efficiency, while preserving adequate model performance. The proposed model is assessed using two publicly accessible datasets, FigShare and Kaggle, and is cross-validated using the BraTS2019 and BraTS2020 datasets, demonstrating high accuracy and explainability. Notably, GATransformer generates interpretable attention maps, visually highlighting tumor regions to aid clinical understanding in medical imaging.

Keywords: graph attention network; Transformer; brain tumor detection; cross-dataset validation; attention maps; explainability

Academic Editor: Maryam Ravan

Received: 20 January 2025
Revised: 31 January 2025
Accepted: 4 February 2025
Published: 6 February 2025

Citation: Tehsin, S.; Nasir, I.M.; Damaševičius, R. GATransformer: A Graph Attention Network-Based Transformer Model to Generate Explainable Attentions for Brain Tumor Detection. *Algorithms* **2025**, *18*, 89. https://doi.org/10.3390/a18020089

1. Introduction

Brain tumor detection heavily depends on medical imaging as well as classification to determine proper treatment planning and prognosis for patients. Combined with other computational methodologies, deep learning has driven significant innovations in analyzing medical images over time. Existing brain tumor detection techniques utilize manually engineered feature extraction and shallow learning algorithms that might not correctly identify complicated patterns in magnetic resonance imaging (MRI). This form of imaging is widely used in neuroimaging to help predict neurological disorders [1]. The latest studies focused on using MRI images to autonomously classify various kinds of tumors in the brain [2]. MRI scan characteristics were extracted using classic ML-based classifiers such as KNN, SVM, and RF [3–5]. Deep learning (DL) models like CNNs and DNNs fared higher in image recognition tests [6,7]. Therefore, deep learning-based classifications are utilized to effectively identify brain tumors via MRI imaging.

Recent advancements in deep learning have enhanced the interpretation of medical images, particularly the implementation of Transformer architectures. Initially designed for natural language processing, it extracts relevant features from high dimensional data while capturing spatial relationships. Despite its development, there are still serious problems

and restrictions regarding brain tumor diagnosis and classification. Conventional deep learning techniques and convolutional neural networks (CNNs) may face obstacles such as the interpretability of predictions, class imbalance, and insufficient information availability. The optimization of hyperparameters in deep learning models is a complex process that frequently necessitates a significant number of computational resources and expertise. This study proposes an innovative strategy for identifying and classifying brain tumors, overcoming existing drawbacks.

These deep learning frameworks demonstrate inductive biases, complexity, extended training times, and significant computational capacities [8]. More enormous databases and data augmentation are required to ensure accurate classification results in the previously mentioned methods [9]. Transformers have been developed to tackle the above-mentioned problems in image processing. The adaptive Transformer architecture, including the visual Transformer, outperformed CNN classifiers and had fewer inductive biases than pre-trained techniques [10]. Transformer global patch-based learning reduces inductive biases and captures important MRI scan properties better than CNN models [11]. Transformer models' multi-head self-attention portion enhances feature engineering by targeting near-tumor regions, outperforming ordinary versions. However, training the Transformer model requires a lot of data and processing power.

Transformer-based models, have enhanced medical image classification [12], including brain tumor diagnosis. These models are accurate but typically operate as black boxes, raising issues about explainability and trustworthiness in essential applications like healthcare. Medical decision making requires human monitoring for safety, reliability, and ethics. Medical AI has largely adopted the Human-in-the-Loop (HITL) [13,14] strategy, which combines automated predictions with expert validation. This method works in radiology, pathology, and clinical diagnostics, where AI aids decision-making but the doctor makes the final call. In this study, we use self-attention techniques for explainability but acknowledge the need for human interaction in understanding AI-driven diagnoses and assuring clinical usability.

This article aims to improve brain tumor classification models' accuracy, efficiency, and interpretability by incorporating Transformer structures and a self-attention graph attention network (GAT). The proposed method enhances the clinical process by providing more accurate and dependable means of diagnosing brain tumors, resulting in improved patient outcomes and a deeper comprehension of neurological disorders. The proposed model employs the Transformer and an attention mechanism, GAT, to identify and preserve the critical channels within the neural network. The channel attention module analyzes the interaction between the weights and the channels, extracting deeper features to improve the representation power. This combination results in a significant increase in computation efficiency and a substantial reduction in model size while maintaining satisfactory performance. Along with the classification accuracy, this work focuses on generating attention maps where the cancerous parts are highlighted. Cross-dataset validation is also carried out, where the proposed model is trained on separate datasets (i.e., FigShare and Kaggle) while tested on unseen datasets (i.e., BraTS2019 and BraTs2020).

This paper is structured in the following order: Section 2 examines recent works linked to the presented work, Section 3 gives a detailed explanation of the suggested technique, Section 4 analyzes the outcomes and performance of the model that has been created, and Section 5 outlines the work's conclusion.

2. Literature Review

Early detection and treatment for tumors in the brain are necessary for positive results. Detecting brain tumors using MRI and CT scans may be difficult because of higher

complexity, more significant computation needs, and elevated false negative rates. An FT-ViT framework was developed for recognizing brain tumors via MRI scans [15]. Techniques include feature processing, feature engineering, patch processing, and fine-tuning. Based on trial results, the CE-MRI database-based model had 98.12% accuracy. However, this method requires a lot of computer power. The authors of [16] applied models trained with deep learning to develop an autonomous system for detecting brain tumors. This approach integrates the features of CNN with a computer-aided diagnostic tool for effective image processing. A transfer-learning-based modified classifier is applied to distinguish tumor kinds. This study used the public brain tumor dataset for verification. Framework sensitivity was 98.77%, accuracy 98.86%, and F-measure 98.71%. The system is lightweight and computationally viable, although overfitting constraints limit its usefulness.

The authors of [17] developed a multiclass categorization strategy for predicting different kinds of brain tumors. The effectiveness of the randomly created TL framework was evaluated against pre-trained models like VGG16/19, ResNet50, Xception, and InceptionV3 using a 3264-image brain dataset. The deployment achieved 97.12% and 94.12% tumor categorization accuracy for binary and multiclass analysis. This accurate classification system helps radiologists make plausible conclusions. However, this causes many false negatives and positives. Transformer models are popular in computer vision due to their low inductive biases and processing requirements. A ViT-based DNN framework was developed for MRI-based brain tumor detection [18]. Pre-trained model-based and fine-tuned ViT skills diagnosed gliomas, pituitaries, and meningiomas 100% accurately. This technique scored 98.7% accuracy on a database of 3064 images of brain tumors. However, there are some problems in analyzing the methods of decision making.

A comprehensive MRI brain tumor segmentation and classification approach was introduced, incorporating three-dimensional CNN and a Transformer framework [19]. This approach accurately retrieves both global and local spatial characteristics from photos. The retrieved features are transformed into tokens and transmitted to the Transformer encoder. The model developed surpasses modern models in terms of precision, as validated by the entire comparative assessment. However, this method needs additional training along with computational resources. A CNN-based technique was presented that can precisely segment and identify tumors in the brain from MRI scans [20]. This approach integrates pre-trained CNNs with adaptable optimization techniques to classify brain tumors. The method utilizes a pre-trained CNN as a classification algorithm and an adaptable dynamic sine-cosine fitness gray wolf optimizer for tuning its hyperparameters. Integrating appropriate fitness into the CNN boosts classification accuracy and error rate. However, this approach is ineffective for categorizing multiple kinds of cancers.

U-Net architectures are widely used for medical image segmentation but face limitations when dealing with small tumor targets and fuzzy boundaries. To address these challenges, the study [21] proposed a modified U-Net architecture incorporating: attention mechanisms to guide the network to focus on relevant regions, improving segmentation accuracy, and multiscale feature fusion to enhance the network's ability to process information across multiple scales, improving segmentation efficacy for diverse tumor structures. Although not explicitly addressed, attention mechanisms could improve explainability by highlighting the regions of interest (e.g., tumor areas) that the model focuses on during segmentation.

Explainability is a critical aspect that ensures that clinicians can trust and interpret the predictions of artificial intelligence (AI) models. Various studies have proposed innovative explainability techniques to make AI-based brain MRI segmentation and classification models more interpretable and reliable. Odusami et al. [22] explored explainability in diagnosing Alzheimer's disease using MRI and PET image fusion. The study employed a

heuristic early feature fusion framework combined with a modified ResNet18 model to extract descriptive features from multimodal data. An explainable artificial intelligence (XAI) model was implemented to interpret the predictions, making the results transparent and enhancing confidence in predictions. Tehsin et al. [23] proposed a disease and spatial attention module (DaSAM)-based model for brain tumor detection. The disease attention module (DAM) identifies disease and non-disease regions, while the spatial attention module (SAM) highlights critical image features. By emphasizing key regions and features, these modules enhance model interpretability and provide clinicians with feature maps that explain the decision-making process. Ullah et al. [24] proposed a framework integrating DeepLabV3+ and custom neural networks for brain tumor segmentation and classification. The explainability component utilized local interpretable model-agnostic explanations (LIME) to interpret the model's outputs. LIME generates human-readable visual explanations by identifying features that significantly contribute to the predictions, offering insights into how the model localized and classified brain tumor regions. However, these methods face several limitations. Attention mechanisms, while helpful in focusing on critical regions, often lack transparency in their own workings and may highlight regions without clear medical relevance, leading to potential misinterpretation. Post hoc techniques like LIME and SHAP provide local explanations but are computationally intensive and can be inconsistent when applied to complex models, raising concerns about reliability. Multimodal fusion approaches struggle with the heterogeneity of data sources like MRI and PET, leading to challenges in generating cohesive and interpretable outputs. These challenges underscore the need for more transparent, domain-specific explainability frameworks tailored to the nuances of medical imaging.

3. Proposed Methodology

This article proposes a novel GATransformer method to generate explainable attention for brain tumor detection. The selected datasets are pre-processed using two augmentation techniques: essential augmentation, where images are flipped and rotated, and SMOTE augmentation. The proposed model calculates correlations and relationships among channel weights. The GAT captures dependencies, while the Transformer computes inter-channel correlations across layers. Figure 1 shows the overall architecture of the proposed model.

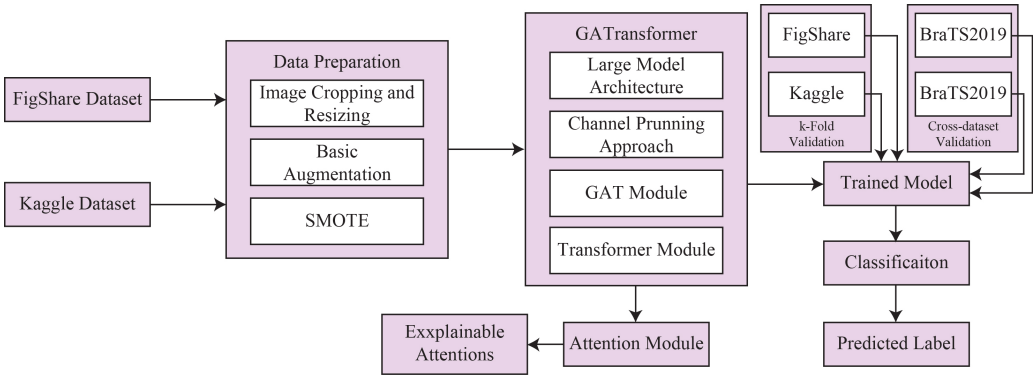

Figure 1. Overall architecture of the proposed model.

3.1. Datasets

The Brain Tumor MRI Dataset (Kaggle Dataset) [25] consists of 7023 MRI image slices, meticulously annotated and classified into four distinct categories. Specifically, it includes 1321 slices corresponding to glioma, 1339 slices representing meningioma, 1595 slices depicting no tumor cases, and 1457 slices of pituitary tumors. This dataset's "no tumor" class images are sourced from the Br35H dataset. Sample images of the Kaggle dataset are shown in Figure 2. The Brain Tumor Dataset (FigShare Dataset) [26] comprises

3064 MRI image slices, each meticulously annotated and categorized into three distinct brain tumor types. Specifically, it contains 708 slices corresponding to meningioma, 1426 slices representing glioma, and 930 slices depicting pituitary tumors. Sample images of the FigShare dataset are shown in Figure 3, whereas Table 1 shows the summary of images in each class of selected datasets.

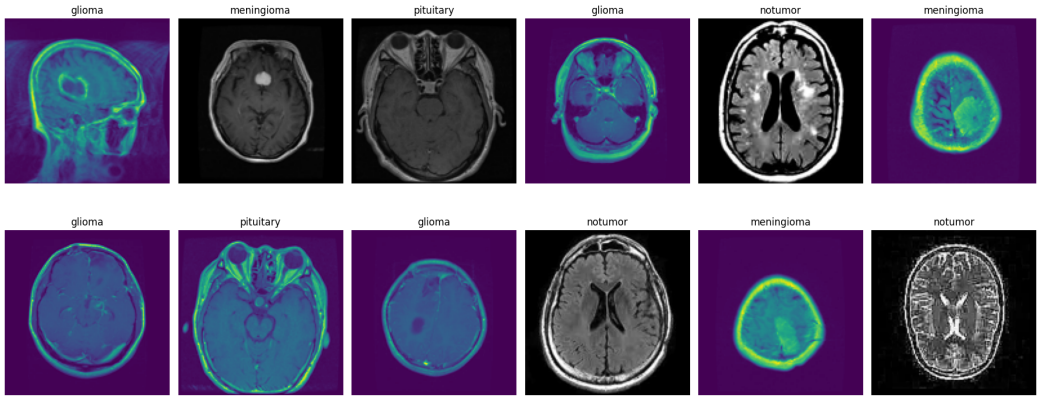

Figure 2. Sample images of the Kaggle dataset.

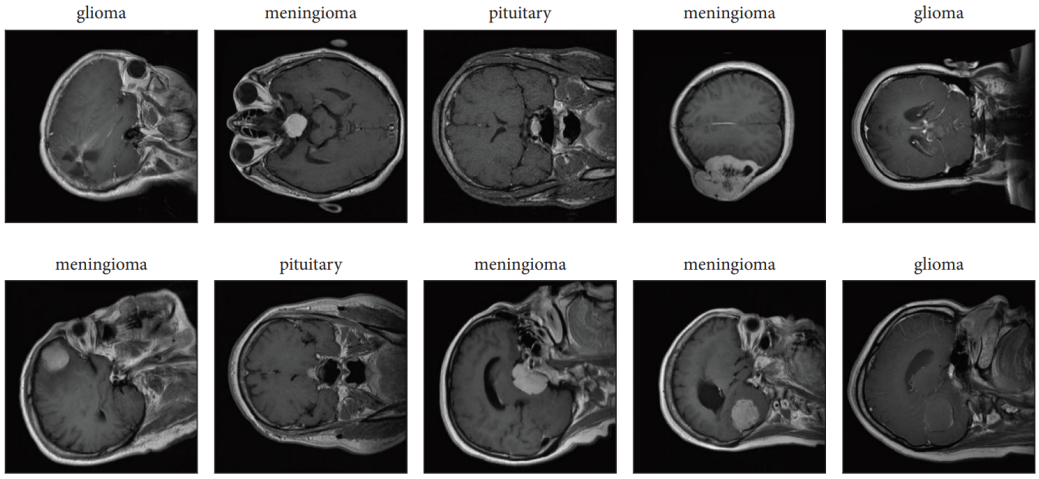

Figure 3. Sample images of the FigShare dataset.

Table 1. Statistics of selected brain tumor datasets.

Dataset	Glioma	Meningioma	Pituitary	No_Tumor	Total
Kaggle [25]	1321	1339	1457	1595	5712
FigShare [26]	1426	708	930	-	3064

The Kaggle dataset includes MRI scans of glioma, meningioma, pituitary tumors, and healthy cases. The FigShare dataset has a variety of MRI scans with the same tumor diagnoses, assuring consistency. These datasets were chosen for their availability, visual diversity, and broad use in past studies, making them reliable standards for deep learning model evaluation. SMOTE was used to improve underrepresented classes and solve data imbalance. Simple augmentations like random rotations and horizontal flipping improved generalization. Our GATransformer model was compatible with datasets pre-processed to normalize image sizes. Both datasets have been extensively used in peer-reviewed medical AI studies, boosting their legitimacy and importance. Public availability facilitates reproducibility and comparability with existing methods, making them ideal for building strong deep learning models in brain tumor classification.

3.1.1. Data Preparation

The aim of this step is to prepare 3 sub-datasets for both selected datasets. The original dataset is kept the same by implementing the cropping strategy and resizing operations as subset 1. Subset 2 is generated by employing the basic augmentation strategy on subset 1, while the third subset is generated by employing the SMOTE strategy on subset 1. Both augmentation strategies are explained in the following sections.

3.1.2. Image Cropping and Resizing

The selected MRI datasets predominantly exhibit undesirable gaps and areas, hence diminishing classification performance. Consequently, cropping images to eliminate extraneous regions is crucial in the image collection. The cropping technique outlined in [27], which relies on extreme points computation, is utilized. Figure 4 illustrates the approach of cropping through extreme point computation. Initially, we provide the original MR images for pre-processing. Secondly, we apply thresholding to the MR images to generate binary images.

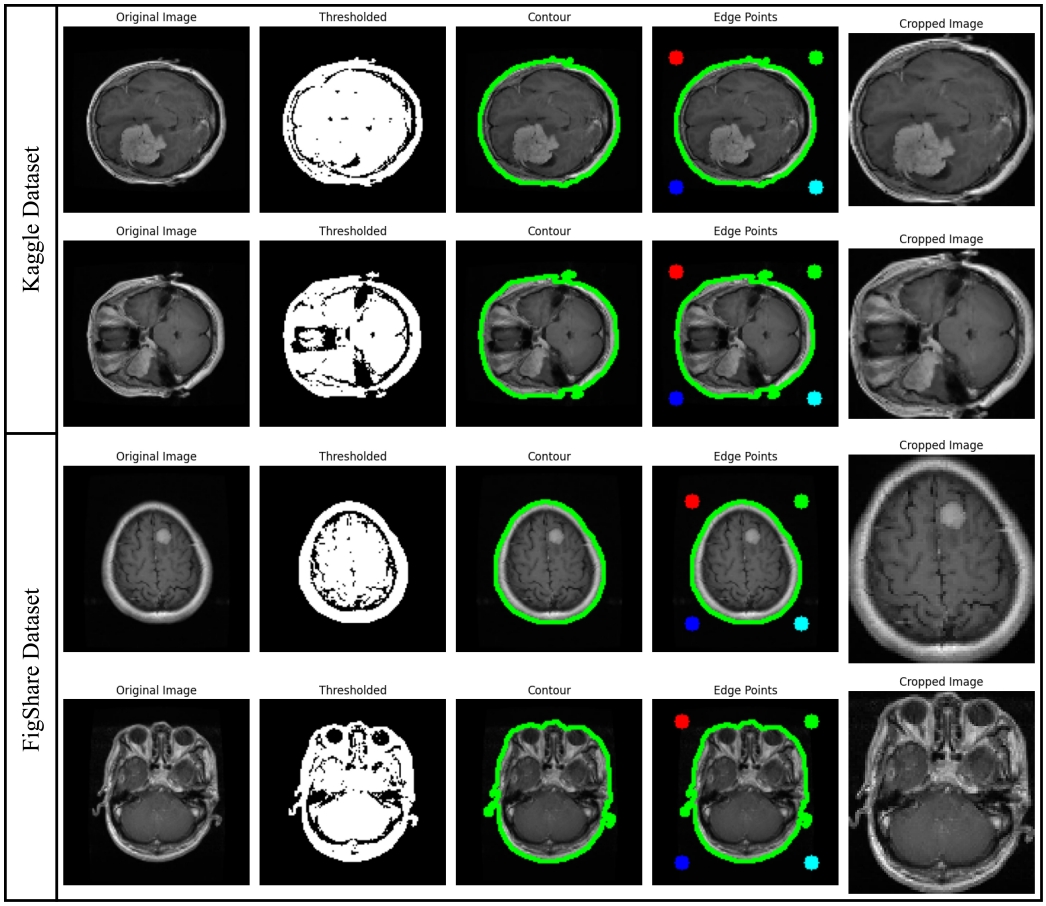

Figure 4. Image cropping using extreme point calculation in two steps.

We performed dilatation and erosion to eliminate noise from the images. Subsequently, we discerned the most prominent contour from the threshold images and continued to pinpoint the four extreme points (most superior, most inferior, rightmost, and leftmost) of the photographs. Finally, we crop the image utilizing contour data and extreme points. The tumor photos are scaled using bicubic interpolation. The primary rationale for selecting bicubic interpolation is its ability to generate a smoother curve compared to alternative approaches like bilinear interpolation, making it more suitable for MR images, which often exhibit significant edge noise. Due to the varying width, height, and sizes of the MR images

in the selected dataset, it is advisable to scale them to uniform dimensions for optimal results. In this study, the MR images are downsized to $250 \times 250 \times 3$ pixels, as the suggested model requires this input dimension. Both datasets are shrunk to the input dimensions of the employed pre-trained CNN models during implementation.

3.1.3. Basic Augmentation

Augmentation improves model learning on most MRI datasets due to their small size and high imbalance. Image augmentation modifies images from the existing dataset to create new datasets for neural network analysis. These augmentations generally change the image's scale, orientation, position, etc. According to sources, the model's categorization accuracy can be enhanced by propagating current data rather than collecting new data. Our basic augmentations created different training sets using two methodologies. A simple method is randomly rotating the input image zero or more times and flipping each rotated output horizontally. For random rotation, each input image was rotated by a random angle ranging from 0° to 360°, guaranteeing that the model remains invariant to various orientations of brain tumors. And for the horizontal flipping, each rotated image was randomly flipped along the horizontal axis, increasing data diversity without altering the anatomical structure of the brain.

3.1.4. SMOTE

The over-sampling method, which involves augmenting the minority class by generating synthetic images from the original minority class images, influenced traditional SMOTE techniques [28] and standard machine learning data augmentation methods like rotation and mirroring. Synthetic examples are generated by operating directly in "image space" rather than "feature space". The minority cases are over-sampled when Gaussian noise is randomly introduced to "n" instances selected from minority cases. The magnitude "n" is ascertained according to the extent to which over-sampling is necessary.

In [25], which consists of 1321 glioma, 1339 meningioma, 1457 pituitary, and 1595 no tumor cases, leave-one-out cross-validation requires balancing the training set by creating or removing synthetic cases. When a glioma case is left out, 19 synthetic meningioma cases, 137 synthetic pituitary cases, and 275 synthetic no-tumor cases must be removed to balance the dataset at 1320 cases per class. Similarly, if a meningioma case is left out, 17 synthetic glioma cases, 136 synthetic pituitary cases, and 274 synthetic no-tumor cases are removed to achieve the same balance. When a pituitary case is excluded, 18 synthetic glioma cases, 18 synthetic meningioma cases, and 274 synthetic no-tumor cases are removed. Finally, if a no-tumor case is left out, 18 synthetic glioma cases, 18 synthetic meningioma cases, and 136 synthetic pituitary cases are removed to balance the dataset at 1321 cases per class. For this work's experiments, 1320 cases per class are kept on this dataset.

In [26], which contains 1426 glioma, 708 meningioma, and 930 pituitary cases, leave-one-out cross-validation similarly requires balancing the training set. If a glioma case is left out, 717 synthetic meningioma cases and 222 synthetic pituitary cases must be removed to balance the dataset at 708 cases per class. When a meningioma case is excluded, 719 synthetic glioma cases and 223 synthetic pituitary cases are cleared. Finally, if a pituitary case is left out, 718 synthetic glioma cases and 221 synthetic meningioma cases are removed to achieve balance. This process ensures that, in each cross-validation iteration, the training set remains balanced across all classes after leaving one instance out. For this work's experiments, 708 cases per class are kept on this dataset.

Synthetic images are produced by the subsequent method: Select m minority examples at random. Compute the standard deviation of the m images. Gaussian noise with a mean of R and S of the standard deviation of the m selected minority-case images is applied to

each of the m photos to produce m additional synthetic minority images. Our approach presently utilizes $R = 0$ and $S = 1$ configurations.

Figure 5a presents an example of an original tumor image, whereas Figure 5b displays the matching synthetic image. Figure 5c illustrates the binary difference map between the original and artificial images. A pixel value 0 in the difference map indicates that no change occurred after adding. A pixel value of 1 indicates that the pixel's intensity in the synthetic image changed due to the added Gaussian noise.

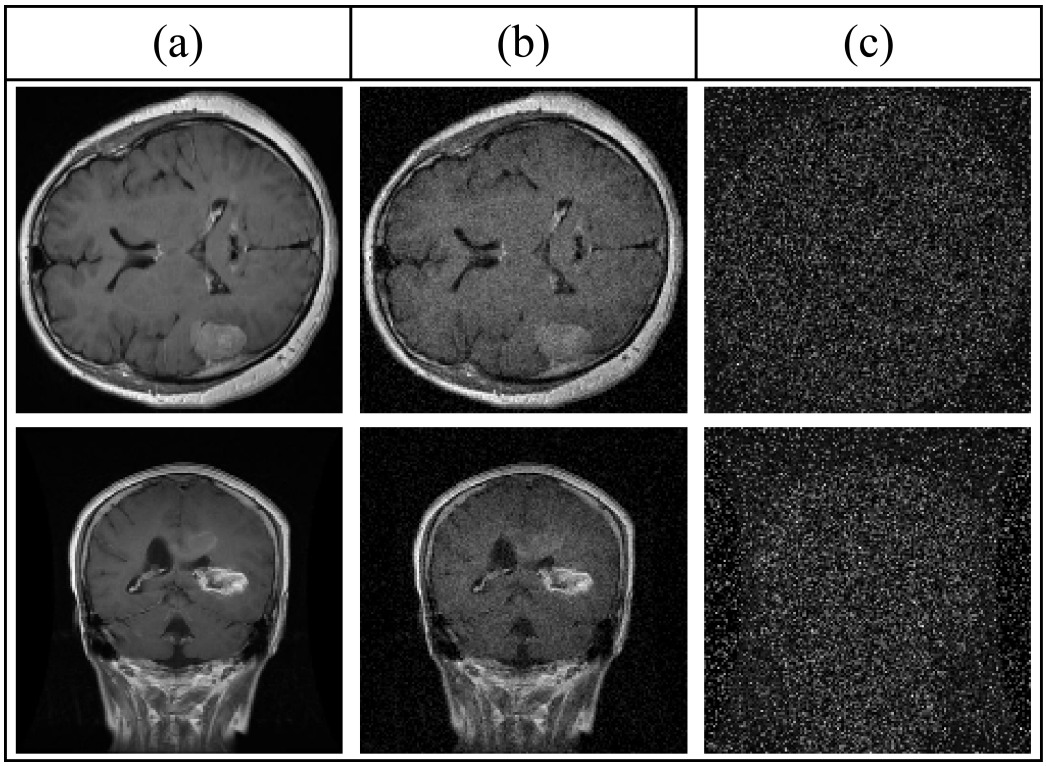

Figure 5. Comparison of the original and SMOTE-generated image; (**a**) original input image; (**b**) matching synthetic image; and (**c**) binary difference map between the original and artificial images.

3.2. Proposed GATransformer

The GATransformer model is a combination of a GAT model [29] and a Transformer [30]. It is designed to reduce the intricacy of computation and architecture in neural architectures without sacrificing performance. GATransformer is employed to learn how to enhance and retain critical channels, assigning importance scores during training to facilitate channel pruning by building upon the attention mechanism. The proposed model calculates correlations and relationships among channel weights. The GAT captures dependencies, while the Transformer computes inter-channel correlations across layers. A feed-forward network processes the output of the multi-head attention network to extract deeper features, while a residual module prevents degradation from the network depth. This combination enhances feature representation, enabling effective channel pruning with minimal performance loss.

3.2.1. Large Model Architecture (LMA)

The LMA consists of large-base models with three pre-trained CNN models and a channel attention module designed to mitigate the excessive memory capacity of large networks. The channel attention module, comprising a GAT and Transformer, employs a two-stage attention mechanism. For this work, a set of three pre-trained CNN models is finalized among selected five pre-trained CNN models including InceptionV3 (In) [31], DenseNet121 (De) [32], EfficientNet-B0 (E0) [33], DarkNet53 (Da) [34], and MobileNetV2

(Mo) [35]. The architecture of LMA with channel pruning to obtain a pruned model is shown in Figure 6.

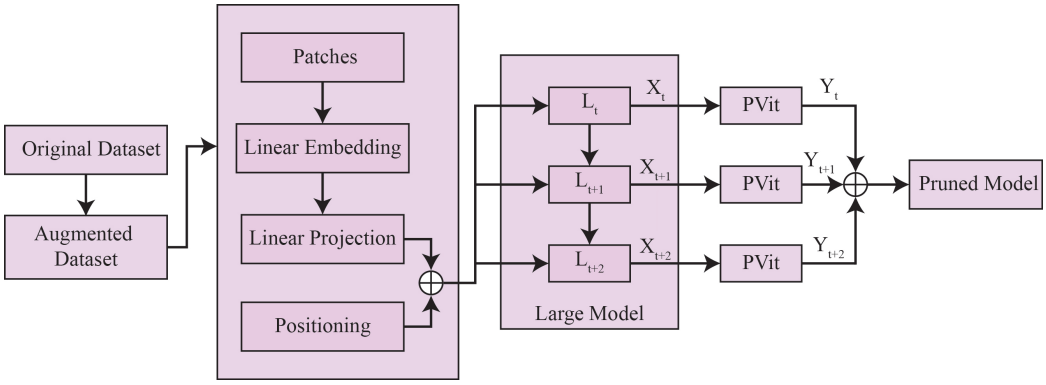

Figure 6. Architecture of LMA with channel pruning to obtain a pruned model.

LMA improves the brain tumor classification feature extraction and representation learning. A model that can capture fine-grained information while keeping global contextual awareness is needed for MRI images' complex spatial and structural patterns. The LMA preserves spatial relationships in MRI images using patch-based embeddings, linear projections, and positional encoding. In addition, the channel pruning approach eliminates redundant computations, making the network more efficient without affecting accuracy despite the high model size. With PViT (pruned vision Transformer), model complexity and computational feasibility are balanced, decreasing redundancy and preserving good feature representation. This method preserves the model's large-scale architecture benefits while optimizing resource use for practical deployment. The model captures detailed tumor characteristics, improves classification performance, and balances model size and efficiency using LMA, making it ideal for high-performance and resource-constrained environments.

The GAT module processes channel weights as nodes to capture intra-layer relationships and assigns attention scores, identifying significant channels within individual layers. The Transformer module analyzes inter-layer relationships using a multi-head attention mechanism, enabling cross-layer channel attention effects. This progressive channel pruning strategy evaluates layer-by-layer dependencies, gradually removing redundant channels across the entire network while maintaining performance. The pruning process is optimized by quantifying discrepancies using mean squared error (MSE) loss. Compared to traditional manual or single-layer pruning methods, this approach incorporates local (GAT) and global (Transformer) attention mechanisms, achieving efficient model compression without significant performance loss. This strategy reduces model complexity while retaining critical network functionalities.

3.2.2. Channel Pruning Approach

The proposed progressive channel pruning technique efficiently prunes superfluous channels while retaining model performance using unique ID assignment and attention mechanisms. Each layer's channel weights in the overall model are flattened into a tensor T to identify each weight as a channel weight node with an ID. This ID-based approach simplifies attention score calculation and channel identification during pruning. The GAT module processes the flattened tensor T to compute the dependencies among channel weight nodes, resulting in the attention scores. These scores capture the intra-layer relationships between channels. Subsequently, the multi-head attention (MHA) mechanism evaluates cross-layer correlations using the GAT output. The MHA output is refined through a feed-forward network (FFN) to extract deeper features. Finally, the production of the MHA is linearly transformed, producing the pruned model. This strategy leverages attention mechanisms at

intra-layer and inter-layer scales, combining the GAT module for local dependency analysis and the Transformer module for global correlation.

3.2.3. GAT Module

This study employs the graph attention network (GAT) to compute attention scores among channel weight nodes and eliminate unnecessary channels, facilitating efficient channel pruning. GAT is a neural network architecture for graph-structured data, which learns node representations by considering structural linkages and nearby node attributes. Each channel weight node has a feature vector describing its features in channel pruning. The attention mechanism in GAT uses a learnable weight matrix to calculate similarity scores between the feature vectors of neighboring nodes. These similarity scores are normalized into attention coefficients, which determine the influence of each neighboring node on the target node. The final representation of a node is derived by combining the weighted sum of its neighbors' features, using the attention coefficients, with its feature vector.

In this study, GAT identifies meaningful channels across the network by calculating attention scores for each channel weight node. These scores rank the importance of channels, enabling the pruning of less significant ones. This progressive pruning method processes layers sequentially, from the first to the last, ensuring that channel importance is consistently evaluated and optimized. By exploring intra-layer and inter-layer relationships, the GAT module enhances the pruning process, improving the model's efficiency while retaining critical channels that contribute to performance. The flattened channel weights from each convolutional layer serve as the input to the GAT module, as illustrated in Figure 7. Through this approach, GAT enables precise and effective channel pruning for large neural networks.

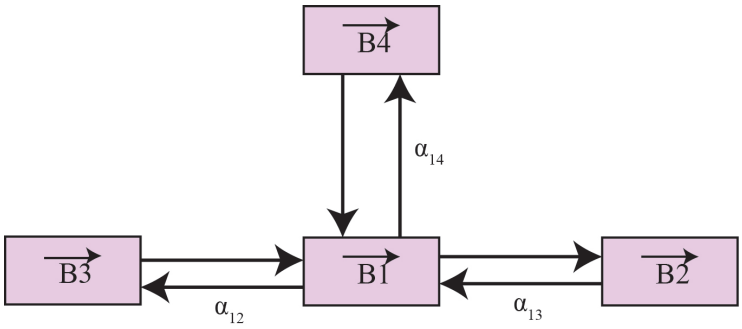

Figure 7. Input to GAT module through flattened channel weights.

We will explain the only graph attentional layer employed in our studies across all GAT topologies. Our attentional arrangement closely follows [36] but it is agnostic to the attention mechanism. Our layer receives node characteristics, $\xi = \{\xi_6, \ldots, \xi_y\}, \quad \xi_x \in \mathbb{R}^d$, where y is the number of nodes and d is their features. The layer generates additional node features (of perhaps varied cardinality d') as $\xi' = \{\xi'_1, \ldots, \xi'_y\}, \quad \xi'_x \in \mathbb{R}^d$ as its output.

At least one learnable linear transformation is needed to turn input features into higher-level features expressively. First, a weight matrix-parameterized shared linear transformation $\omega \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is applied to each node. A shared attentional mechanism $\alpha : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$ computes attention coefficients to perform self-attention on nodes.

$$c_{j,k} = \alpha(\omega\xi_j, \omega\xi_k) \tag{1}$$

This equation shows how vital the node's attributes are to node j. In the most general model, any node can attend any other node, eliminating all structural information. To incorporate graph structure into the mechanism, we use masked attention to compute $c_{j,k}$ solely

for nodes $k \in \eta_j$, where η_j is a neighborhood of node j in the network. All our trials will make them the first-order neighbors of j, including j. To ensure coefficient comparability across nodes, we normalize them using the softmax function ρ for all k choices:

$$\varsigma_{j,k} = \rho_k(c_{j,k}) = \frac{\exp(c_{j,k})}{\sum_{i \in \eta_j} \exp(c_{j,k})} \tag{2}$$

Our experiments use a single-layer feedforward neural network (a) with a weight vector $\in \mathbb{R}^{2d'}$ and LeakyReLU nonlinearity L (with negative input slope $\varsigma = 0.2$) for the attention mechanism. When fully enlarged, the attention mechanism's coefficients, as shown in Figure 8a, become:

$$\varsigma_{j,k} = \frac{\exp\left(L\left(w^T[\omega\xi_j \oplus \omega\xi_k]\right)\right)}{\sum_{i \in \eta_j} \exp\left(L\left(w^T[\omega\xi_j \oplus \omega\xi_k]\right)\right)} \tag{3}$$

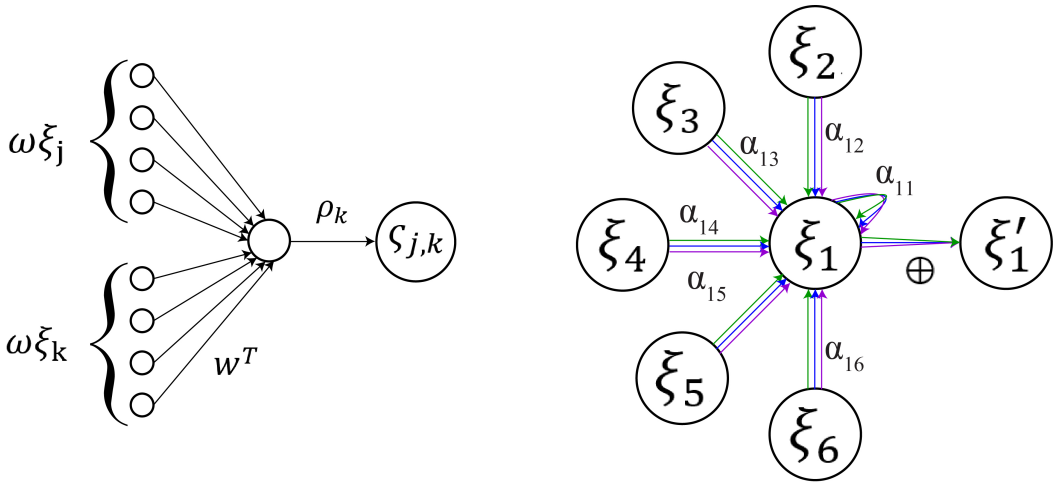

Figure 8. (**a**) The proposed attention mechanism; (**b**) Overall process of multi-head self-attention with three heads by one node.

$\oplus$ represents the concatenation operation, and T represents transposition. To generate output features for each node, the normalized attention coefficients are utilized to compute a linear combination of features corresponding to them, optionally with a nonlinearity σ:

$$\xi'_j = \sigma\left(\sum_{k \in \eta_j} \varsigma_{j,k}\, \omega\xi_j\right) \tag{4}$$

Following the method adopted by [30], we discovered that multi-head attention stabilizes self-attention learning. B separate attention mechanisms convert Equation (4) and concatenate their features to produce the following output feature representation:

$$\xi'_j = \prod_{b=1}^{B} \sigma\left(\sum_{k \in \eta_j} \varsigma^b_{j,k}\, \omega^b \xi_j\right) \tag{5}$$

$\prod$ denotes concatenation function, $\varsigma^b_{j,k}$ are normalized attention coefficients from the bth attention mechanism, and ω^b is the weight matrix of the input linear transformation. The final output ξ' will have Bd' characteristics instead of d'. Concatenation is no longer sensible if we execute multi-head attention on the network's final (prediction) layer. Instead, we use averaging and postpone applying the final nonlinearity (typically a softmax or

logistic sigmoid for classification problems), calculated as the following equation. In contrast, Figure 8b shows multi-head graph attentional layer aggregation.

$$\xi_j' = \sigma\left(\frac{1}{B}\sum_{b=1}^{B}\sum_{k\in\eta_j}\varsigma_{j,k}^{b}\,\omega^{b}\xi_j\right) \tag{6}$$

The GAT attention mechanism determines channel weight associations. The network prioritizes each node's most essential channel neighbors above all channel neighbors. This approach can consider the extent to which channels affect one another and the connections between channels across several layers.

3.2.4. Transformer Module

The limitations of exclusively relying on attention weights from GAT for channel pruning are identified in this study, which also recognizes the necessity of a more comprehensive methodology. The GAT's primary emphasis is on the attention weights of individual layers, and it does not establish connections between them. It is widely acknowledged that the CNN layer at the preceding stage substantially impacts the results of the succeeding CNN layer. This article recommends integrating the GAT and Transformer models to optimize the efficacy of channel pruning and improve interlayer connections. The Transformer paradigm's multi-head attention mechanism is essential to this goal. The multi-head attention mechanism, initially developed for natural language processing, has been applied to several fields.

The Transformer module has four parts, as shown in Figure 9. The Transformer encoder block sends GAT module output to the classification and attention blocks for classification and parallel attention map generation. The attention restructuring block comprises three mechanisms. The initial step is to divide the attention extracted from GAT into a series of identical segments. The Transformer uses the segment count to compute the input sequence length. A trainable linear embedding compresses the patches into a sequence of identical tokens via a linear projection. This is necessary because Transformers require a consistent latent vector size across all layers. Finally, the embedded tokens are positionally embedded to retain the patch position. Standard, learnable 1D position embeddings are used. The Transformer encoder block will receive the output of these operations as input.

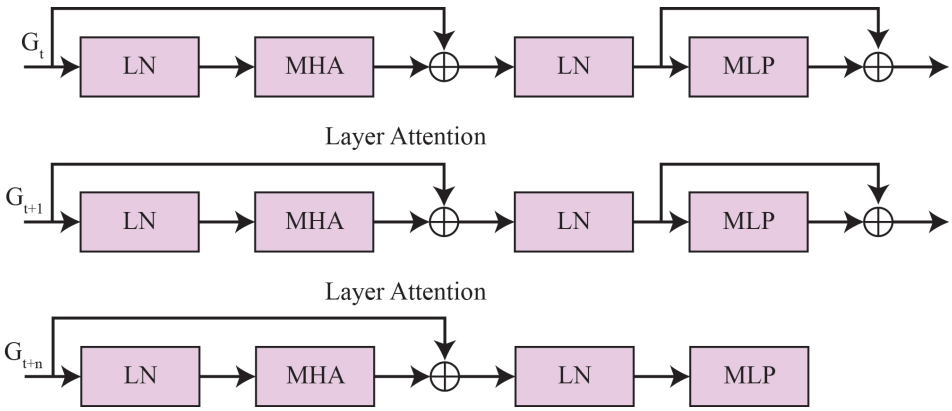

Figure 9. Architecture of the proposed Transformer module.

The Transformer encoder has L iterative normalization layers with two sublayers. The first sublayer is multi-head self-attention. Sublayer two is a complete feed-forward network. All sublayers receive normalization [37] and residual connection [38]. Each sublayer outputs $I + SL(Norm(I))$, where SL is its function over normalized input I. First,

investigate an attention function to understand multi-head self-attention. An attention function maps a query and key–value pairs to an output. People call attention to scaling dot-product attention. The attention function receives vectors with a query and key-value. A compatibility function evaluates the query-key connection to determine the weights of the items in the weighted sum. Attention, or self-attention, associates points in an input sequence to represent it.

Consider the inputs Q, K, and V for the queries, keys, and values of dimensionality dim. The weight matrices ω_Q, ω_K, and ω_V are multiplied with the embeddings of the input vector I acquired through training. The final calculation involves the combination of Q and K, which are then divided by the square root of the dimensionality of the key vectors. The softmax function defines normalization such that each output vector sums to one. The attention of the input is multiplied by V.

$$Attn(K, Q, V) = \rho\left(\frac{QK^T}{\sqrt{dim_K}}\right)V \tag{7}$$

Multi-head self-attention uses a concurrent attention function to focus on information from different representation subspaces at different locations. The operation results are concatenated and converted into linear units of the expected dimension:

$$mHSA(K, Q, V) = \bigoplus(h_1, \ldots, h_g)\omega^{out} \tag{8}$$

$$h_i = Attn(K\omega_K^i, Q\omega_Q^i, V\omega_V^i) \tag{9}$$

The projections are parameter matrices $\omega_K^i \in \mathbb{R}^{dim_m \times dim_K}$, $\omega_Q^i \in \mathbb{R}^{dim_m \times dim_Q}$, and $\omega_V^i \in \mathbb{R}^{dim_m \times dim_V}$. The GAT module output dimension is represented by dim_m. Additional output processing is possible using the MLP sublayer. It has two wholly connected layers and a GeLU nonlinearity layer [39].

3.2.5. Classification Module

The collection of T Transformer encoders produces a sufficient range of features for brain tumor categorization. Tumor classification is the block's purpose. It has five layers: flatten, normalization, dense, normalization, and softmax. The flattening layer flattens the Transformer encoder output into a one-dimensional array. The flattened layer output is batch normalized and fed into the dense layer to recenter and rescale the input distribution to avoid vanishing or bursting gradients. The thick layer of a deep neural network uses a GeLU activation function, proposed in [40], which outperforms ReLU. The output of the dense layer is subjected to a second batch normalization. The final classification outcome is achieved using a softmax layer.

3.2.6. Attention Module

The attention weights produced during the self-attention mechanism are extracted via Equation (7). Self-attention calculates token weights based on their relevance in the input sequence. These weights are computed by dotting query and key vectors, scaling, masking, and softmax. After this, TensorFlow enables access to these attention weights through model outputs and forwards these attention weights to attention layers to capture the attention matrices during inference. These matrices are then normalized and visualized as heatmaps to interpret the interactions between tokens, providing insights into the model's decision-making process.

4. Experimental Results

In this section, we delve into the proposed performance analysis model employed in classifying brain tumors in depth. Through rigorous evaluation and assessment, we aim to uncover the strengths and weaknesses of the proposed model, ultimately identifying the most effective approach for accurate brain tumor classification.

4.1. Experimental Setup

The experiments were conducted on a robust computing setup equipped with Ryzen 7 7700 with MSI PRO B650M-A DDR5 motherboard with a core i7, 12th generation processor having 12 cores and 20 threads, 32 GB RAM, M.2 1 TB NVME SSD, and an MSI GeForce RTX 4060 8 GB Ventus 2X Black 8 GB graphics card. The experimental framework was established within a Jupyter Notebook environment, which offers a flexible and interactive platform for conducting analyses. The proposed methodology uses Python with libraries like Scikit-learn, Keras, TensorFlow, NumPy, Pandas Seaborn, and Matplotlib. The proposed model implemented leave-one-out cross-validation for SMOTE and k-fold cross-validation for model training to ensure the robustness and generalizability of the model. Three sub-datasets were used, including the original dataset, the basic augmented dataset, and the SMOTE augmented dataset. All datasets were split using the 70–15–15 approach for training, testing, and validation. The Adam optimizer with a learning rate of 0.0001 was employed using a batch size of 32 and early stopping. The model was trained on 99 epochs before it was stopped by meeting the early stopping criteria.

4.2. Performance Metrics

The key performance metrics used in our experiment include accuracy (A), precision (P), recall (R), F1-score (F1), and root mean squared error (RMSE) [41]. All these metrics are generally used to evaluate the brain tumor models in state-of-the-art (SOTA) methods. RMSE is the square root of the mean squared error. It is more sensitive to outliers than the mean absolute error (MAE) and is often used in regression analysis. It is calculated as the following equation, where L_c is the actual label, $\hat{L}_c$ is the predicted label, and K denotes the total number of samples in the dataset:

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{c=0}^{C}\left(L_c - \hat{L}_c\right)^2} \tag{10}$$

4.3. Classification Analysis

The proposed model starts with an LMA; in the first experiment, we will evaluate which combination of three pre-trained CNN models performed better on both selected datasets. Table 2 demonstrates the ten late model aggregation performances combined across Kaggle and FigShare datasets, measured by accuracy (A), recall (R), precision (P), F1-score, and RMSE. The results highlight significant variability, with accuracy values ranging from 52.8% to 85.0% on the Kaggle dataset and 56.9% to 85.3% on the FigShare dataset. Among the combinations, IN + DE + MO demonstrates the best performance, achieving an accuracy of 85.0% on Kaggle dataset and 85.3% on FigShare dataset, with precision of 84.2% and 83.4%, recall of 83.8% and 82.0%, and F1-scores of 84.0% and 82.7%. It also recorded the lowest RMSE of 6.6% on the Kaggle dataset and 9.5% on the FigShare dataset. The second-best combination is IN + E0 + DA, which achieves an accuracy of 82.1% on Kaggle dataset and 82.2% on FigShare dataset, with precision rates of 82.7% and 82.1%, recall rates of 82.8% and 81.8%, and F1-scores of 82.1% and 81.9%, along with RMSE values of 14.8% on Kaggle dataset and 10.1% on FigShare dataset. DE + DA + MO ranks third with an accuracy of 81.5% on the Kaggle dataset and 80.0% on FigShare dataset, supported

by precision rates of 81.4% and 81.1%, recall rates of 80.3% and 81.3%, and F1-scores of 82.3% and 81.2%, with RMSE values of 12.8% on Kaggle dataset and 12.4% on FigShare dataset. These top three combinations stand out for their strong classification performance across the datasets.

Table 2. Comparison of pre-trained CNN models to find the best combination for the LMA module on the original dataset. **Bold** shows the best performance; *Italic* shows the second best performance; underlined shows the third best performance.

LMA Combination	Kaggle Dataset					FigShare Dataset				
	A (%)	**P (%)**	**R (%)**	**F1 (%)**	**RMSE (%)**	**A (%)**	**P (%)**	**R (%)**	**F1 (%)**	**RMSE (%)**
In + De + E0	63.6	61.9	62.9	62.4	17.2	61.2	60.9	58.9	59.9	23.0
In + De + Da	74.7	74.8	74.0	74.4	18.8	76.6	73.7	75.4	74.5	14.7
In + De + Mo	**85.0**	**84.2**	**83.8**	**84.0**	**6.6**	**85.3**	**83.4**	**82.0**	**82.7**	**9.5**
In + E0 + Da	*82.1*	*82.7*	*82.8*	*82.1*	*14.8*	*82.2*	*82.1*	*81.8*	*81.9*	*10.1*
In + E0 + Mo	71.5	70.7	71.4	70.0	16.0	69.7	67.7	66.4	67.0	14.3
In + Da + Mo	52.8	51.4	52.4	51.9	15.6	56.9	56.0	55.1	55.5	14.1
De + E0 + Da	72.1	73.8	73.2	72.5	11.1	77.3	74.5	76.0	75.2	13.5
De + E0 + Mo	75.9	75.9	74.6	76.7	18.5	79.7	78.6	77.7	78.1	16.4
De + Da + Mo	81.5	81.4	80.3	82.3	12.8	80.0	81.1	81.3	81.2	12.4
E0 + Da + Mo	70.1	70.3	71.2	69.7	23.0	68.2	65.6	67.5	66.5	15.4

As we have found the best combination of pre-trained CNN models to be integrated into the LMA module, we will test the best combination (In + De + Mo) to evaluate the impact of data augmentation techniques. Table 3 compares the performance of the best combination in LMA on different employed augmentation techniques with the original dataset. During this evaluation, SMOTE demonstrates the most substantial improvements in classification performance, achieving the highest accuracy of 89.3% on the Kaggle dataset and 89.6% on the FigShare dataset. It consistently outperforms in precision, recall, and F1-scores while maintaining the lowest RMSE values of 4.2% on the Kaggle dataset and 6.9% on the FigShare dataset, indicating a significant reduction in prediction errors. Essential augmentation further enhances the results compared to the original dataset, with an accuracy of 87.1% on the Kaggle dataset and 87.5% on the FigShare dataset. It also improves the precision, recall, and F1-scores while reducing the RMSE to 5.5% on the Kaggle dataset and 8.3% on the FigShare dataset. While providing a baseline accuracy of 85.0% on the Kaggle dataset and 85.3% on the FigShare dataset, the original dataset shows higher RMSE values of 6.6% on the Kaggle dataset and 9.5% on the FigShare dataset. These findings emphasize the effectiveness of augmentation techniques like SMOTE and essential augmentation in improving classification outcomes compared to using the original dataset alone.

Now that we have compared the impact of employed augmentation techniques, we will compare the performance of the proposed GATransformer model on both datasets. Table 4 illustrates the performance of the GATransformer model on the Kaggle and FigShare datasets under different augmentation techniques, including the original dataset, essential augmentation, and SMOTE. The results indicate progressive improvement in the classification performance with augmentation techniques. Using the original dataset, the model achieves an accuracy of 92.0% on the Kaggle dataset and 92.2% on the FigShare dataset, supported by precision rates of 91.5% and 91.0%, recall rates of 91.2% and 90.8%, and F1-scores of 91.3% and 90.9%. However, the RMSE values remain relatively higher at 8.0%

on the Kaggle dataset and 7.8% on the FigShare dataset. With essential augmentation, the model demonstrates a notable improvement, achieving an accuracy of 95.0% on the Kaggle dataset and 95.3% on the FigShare dataset. The precision, recall, and F1-scores also increased, reaching 94.6%, 94.4%, and 94.5% on the Kaggle dataset and 94.7%, 94.2%, and 94.4% on the FigShare dataset. RMSE values were reduced to 5.5% on the Kaggle dataset and 5.3% on the FigShare dataset, indicating enhanced prediction stability. The application of SMOTE yields the best results, with the model achieving an accuracy of 99.0% on the Kaggle dataset and 99.2% on the FigShare dataset. Precision, recall, and F1-scores are exceptionally high at 98.7%, 98.5%, and 98.6% on the Kaggle dataset and 98.8%, 98.6%, and 98.7% on the FigShare dataset. The RMSE values are minimized to 2.0% on the Kaggle dataset and 1.8% on the FigShare dataset, showcasing the significant impact of SMOTE in reducing prediction errors and enhancing the overall performance of the GATransformer model.

Table 3. Performance of the best combination in LMA on different employed augmentation techniques.

Dataset	Kaggle Dataset					FigShare Dataset				
	A (%)	**P (%)**	**R (%)**	**F1 (%)**	**RMSE (%)**	**A (%)**	**P (%)**	**R (%)**	**F1 (%)**	**RMSE (%)**
Original dataset	85.0	84.2	83.8	84.0	6.6	85.3	83.4	82.0	82.7	9.5
Basic augmentation	85.0	84.2	83.8	84.0	6.6	85.3	83.4	82.0	82.7	9.5
SMOTE	87.1	86.7	86.3	86.5	5.5	87.5	86.4	86.0	86.2	8.3

Table 4. Performance of the GATransformer model on different employed augmentation techniques.

Dataset	Kaggle Dataset					FigShare Dataset				
	A (%)	**P (%)**	**R (%)**	**F1 (%)**	**RMSE (%)**	**A (%)**	**P (%)**	**R (%)**	**F1 (%)**	**RMSE (%)**
Original dataset	92.0	91.5	91.2	91.3	8.0	92.2	91.0	90.8	90.9	7.8
Basic augmentation	95.0	94.6	94.4	94.5	5.5	95.3	94.7	94.2	94.4	5.3
SMOTE	99.0	98.7	98.5	98.6	2.0	99.2	98.8	98.6	98.7	1.8

Figure 10 displays confusion matrices for two datasets, Kaggle and FigShare, using the proposed model on SMOTE data. For the Kaggle dataset in Figure 10a, the model performs well with Glioma, achieving 1307 correct predictions and minimal misclassifications in the other classes. Meningioma has 1325 true positives, with a few misclassifications into the different courses. Pituitary shows a very high actual positive rate of 1440, and no tumor also performs well with 1574 correct predictions. The confusion matrix is dominated by true positives, which signifies Kaggle's well-performing model. For the FigShare dataset in Figure 10b, the model's performance is similarly good but with slight differences in misclassification. Glioma has 1411 correct predictions, meningioma has 693 correct predictions, and pituitary has 920 true positives, with a few misclassifications to other classes. The misclassifications for no tumor are pretty low. In both datasets, the model struggles slightly with predicting meningioma and pituitary classes accurately, as evidenced by their misclassifications. Still, overall, the model's accuracy is relatively high, as reflected in the overall trends of the confusion matrices. The color intensity in the heatmaps highlights these performance patterns, with darker shades representing higher numbers of predictions. Figure 11 shows the training and validation accuracy and loss on both datasets for the proposed model using the SMOTE dataset.

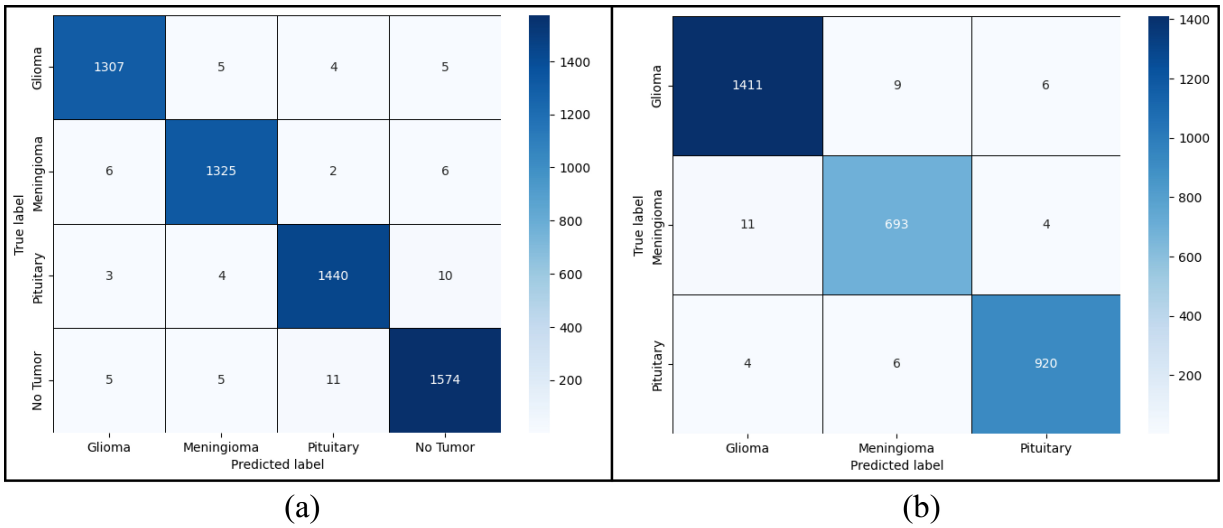

Figure 10. Confusion matrix of GATransformer model on both datasets using SMOTE data; (**a**) confusion matrix for Kaggle dataset and (**b**) confusion matrix for FigShare dataset.

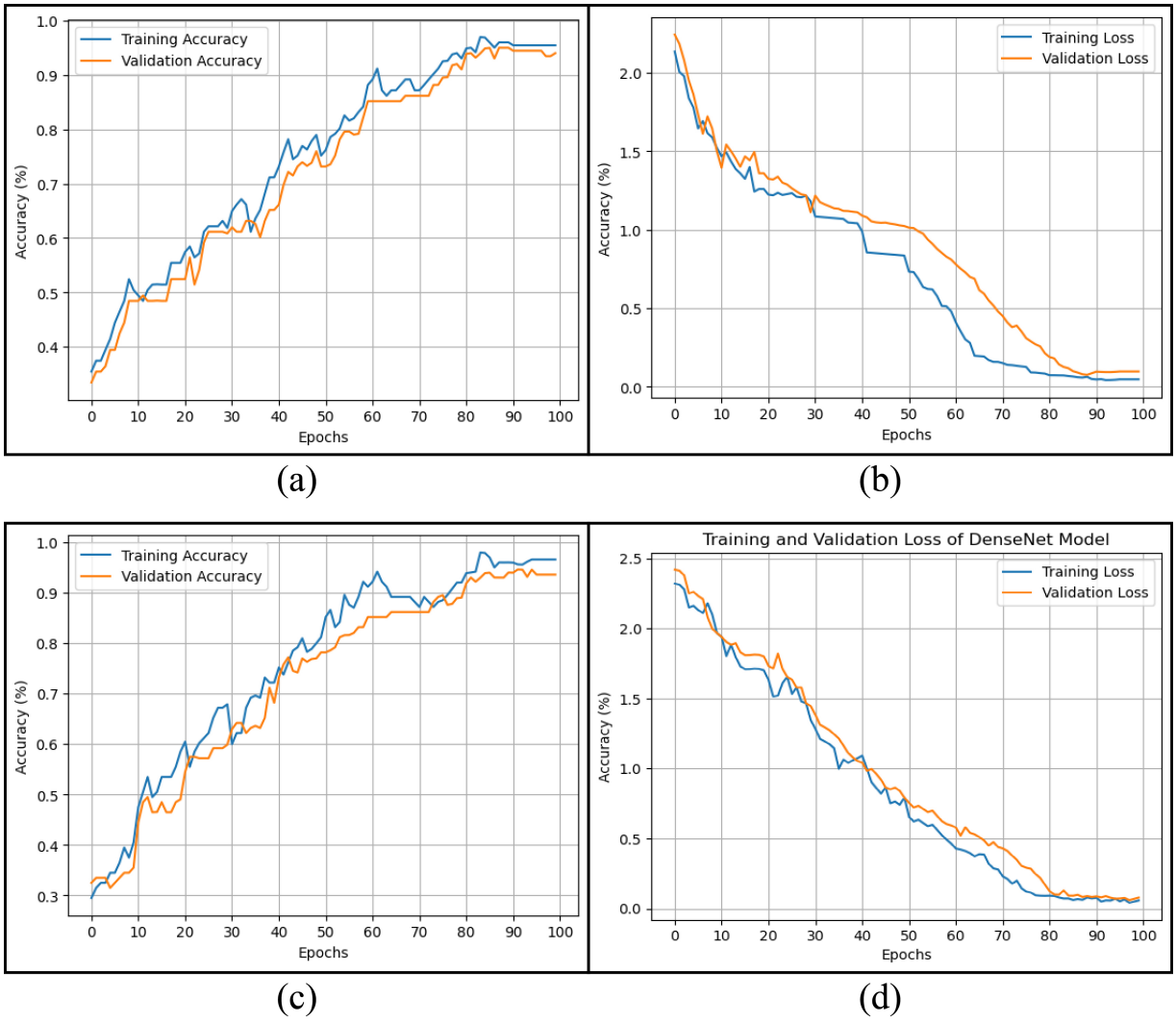

Figure 11. Comparison of the training and validation accuracy and loss of GATransformer model on the SMOTE augmented datasets; (**a**) training and validation accuracy comparison on Kaggle dataset; (**b**) training and validation loss comparison on Kaggle dataset; (**c**) training and validation accuracy comparison on FigShare dataset; and (**d**) training and validation loss comparison on FigShare dataset.

4.4. Attention Analysis

The two different categories of experimentation have undergone attention analysis. The same dataset was used for training and testing in the first set of experiments. Figure 12 shows the proposed model's classification and attention accuracy on the trained and tested Kaggle dataset.

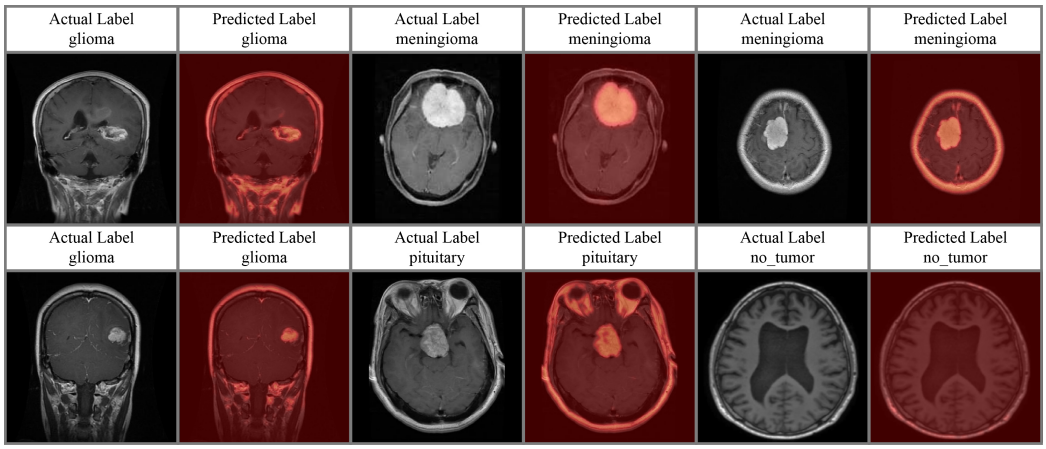

Figure 12. Attention analysis of the proposed model when trained and tested on the Kaggle dataset.

Figure 13 shows the results of the proposed model when trained and tested on the FigShare dataset. The attention shows MRI images in pairs with actual and expected labels for glioma, meningioma, pituitary, and no-tumor cases. These results show that the model can accurately detect and predict tumor locations, as shown by the strong label alignment. This evaluation indicates that the model's resilience and consistency on the dataset it was trained on are sufficient for tumor classification and detection in the same area.

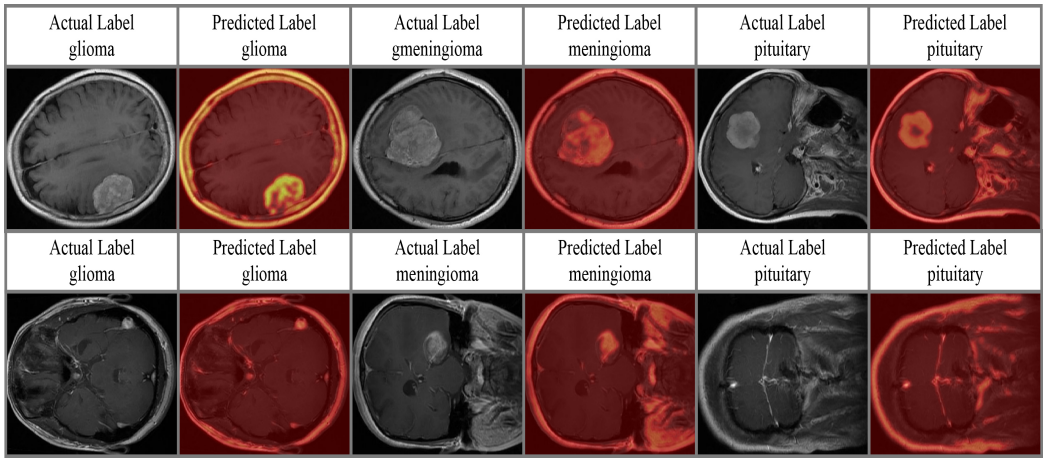

Figure 13. Attention analysis of the proposed model—training and testing dataset: FigShare.

We also examined the impact of cross-dataset validation, where the proposed model is trained on different MRI images while testing on completely unseen data. Three types of attention analysis are carried out in this work. In the first analysis, the proposed model is trained on the FigShare dataset and tested on the Kaggle Dataset. The proposed model's generalization capability across datasets is demonstrated in Figure 14, showing the results of its training on the FigShare dataset and testing on the Kaggle dataset. The figure illustrates pairings of MRI scans, including the actual label and its corresponding computed label. In distinct cases, the three classes of brain tumor—glioma, pituitary, and meningioma—are illustrated, along with their respective actual and predicted segmentation or classification results. Specific instances from the dataset are highlighted by the input image name in the

dataset (e.g., Tr-gl_0165) in each case. The precise overlap between the actual and predicted tumor regions in the images illustrates the model's efficacy in cross-dataset prediction. Despite the domain disparities between the training and testing datasets, this visualization demonstrates the model's ability to handle unseen data during testing.

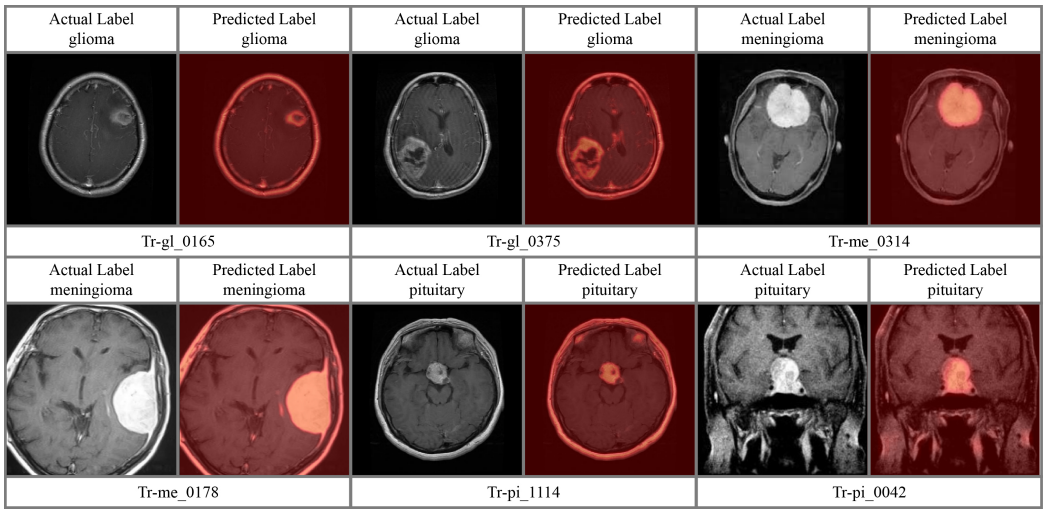

Figure 14. Analysis of the proposed model—training dataset: FigShare; testing dataset: Kaggle.

The proposed model's performance is demonstrated in Figure 15, where the training has been performed on the Kaggle dataset and subsequently testing on the BraTS2019 and BraTS2020 datasets. The MRI image, the ground truth segmentation mask, and the corresponding attention map generated by the proposed model are all presented in three columns. The attention maps emphasize the regions the model identified as relevant for segmentation and classification, demonstrating a significant overlap with the ground truth segmentation. The model accurately segments tumor regions across various test instances in the BraTS2019 and BraTS2020 datasets, as evidenced by the close correspondence between the attention maps and the ground truth masks. The results suggest that the model can accommodate tumor morphologies, sizes, and variations in intensity, despite training on a distinct dataset. This analysis emphasizes the model's capacity to effectively adjust data from various medical image repositories that are not visible.

Medical imaging quality affects the brain tumor classification model performance. Despite deep learning advances, limited resolution, noise, and distortions in MRI scans can conceal tumor features and hamper categorization. Super-resolution methods, especially those based on generative adversarial networks (GANs), can improve image clarity and feature extraction. Recent studies like [42] have shown that AI-driven super-resolution approaches can improve medical imaging resolution while preserving structural integrity. Super-resolution can show finer tumor characteristics by producing high-resolution MRI scans from low-resolution ones, improving categorization. Such solutions in our pipeline could enhance our augmentation efforts and improve GATransformer model feature representation. Our current strategy uses random rotation and flipping; however, super-resolution as a pre-processing step could improve model robustness and classification accuracy. Future research may use GAN-based super-resolution models to enhance MRI images before feeding them into the GATransformer architecture to reduce misclassification due to low picture quality. Super-resolution and deep learning can bridge the gap between AI-based tumor classification and clinically relevant decision support systems, improving diagnostic reliability and model generalization.

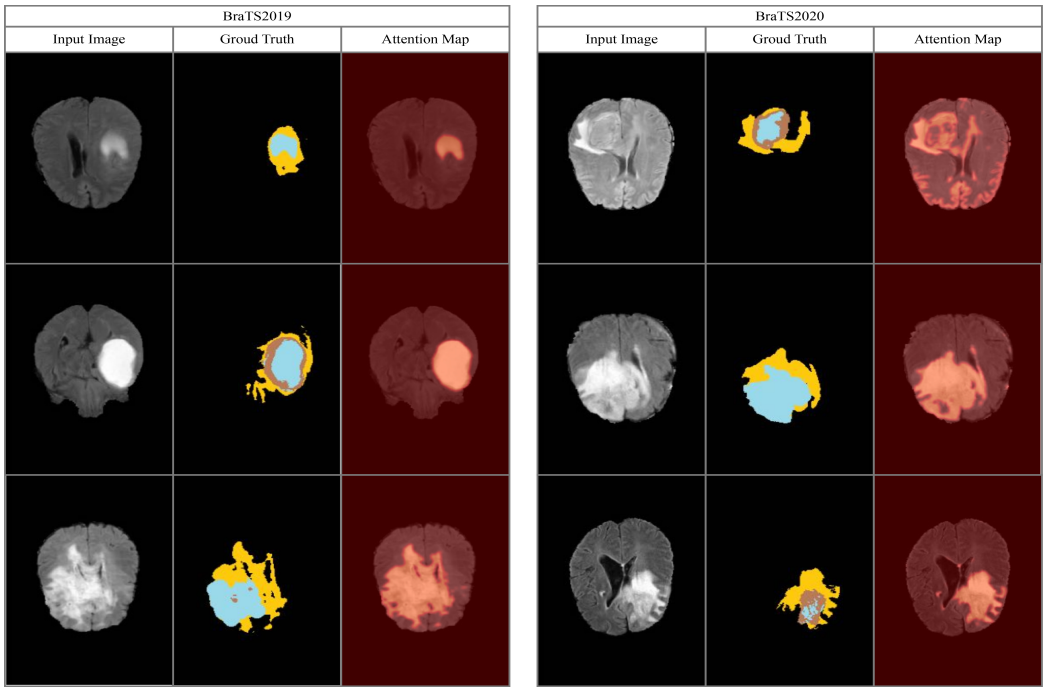

Figure 15. Analysis of the proposed model—training dataset: Kaggle; testing dataset: BraTS2019 and BraTS2020 datasets.

5. Conclusions

In this article, a GAT-based transformer called GATransformer employs the attention mechanism, GAT, and Transformer to detect and sustain patient care neural network channels without medical competence. Channel attention improves model representation by extracting more profound attributes from weight-channel relationships and facilitates channel pruning, while Transformer models can learn complex spatial connections. Integrating these aspects reduces model size and improves computer efficiency while maintaining model performance. The FigShare and Kaggle datasets were used to initially train the proposed model, whereas BraTS2019 and BraTS2020 datasets were employed to cross-validate the proposed model. Ground-truth images from these two datasets were employed to validate the proposed model. The GATransformer model's resource-intensive computational complexity and training time limit its real-time deployment. Even with SMOTE for class balancing, utilizing Kaggle and FigShare datasets increases the overfitting risk. Additionally, model performance depends on high-quality MRI scans, making it sensitive to noise and artifacts. While explainable attention processes are used, deep learning models, especially those using GAT and Transformers, remain "black boxes" and pose interpretability issues. Medical AI explainability is difficult, but our GATransformer model classifies brain tumors well. Healthcare is vital, so we highlight the human-in-the-loop (HITL) strategy to use AI as a decision-support tool rather than replacing doctors. HITL works in radiology and diagnostics, where AI aids but experts approve. Our model improves efficiency and clinical application trust by following this paradigm. More efficient and accurate pruning algorithms and channel pruning in more extensive and more complicated networks should be studied in the future. Channel pruning is promising and may help construct more efficient and powerful deep neural networks.

Author Contributions: Conceptualization, I.M.N.; Methodology, R.D.; Software, S.T.; Validation, S.T., I.M.N. and R.D.; Formal analysis, S.T. and I.M.N.; Investigation, S.T.; Resources, I.M.N.; Writing—original draft, S.T. and I.M.N.; Writing—review and editing, R.D.; Visualization, S.T.; Supervision, R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This study uses the following openly available datasets: Brain Tumor MRI Dataset (Kaggle Dataset) [25], Brain Tumor Dataset (FigShare Dataset) [26], BraTS2019 dataset (https://www.kaggle.com/datasets/debobratachakraborty/brats2019-dataset, accessed on 24 October 2024) and BraTS2020 datasets (https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation, accessed on 24 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shoeibi, A.; Khodatars, M.; Jafari, M.; Ghassemi, N.; Moridian, P.; Alizadehsani, R.; Ling, S.H.; Khosravi, A.; Alinejad-Rokny, H.; Lam, H.K.; et al. Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review. *Inf. Fusion* **2023**, *93*, 85–117. [CrossRef]
2. Ranjbarzadeh, R.; Caputo, A.; Tirkolaee, E.B.; Ghoushchi, S.J.; Bendechache, M. Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. *Comput. Biol. Med.* **2023**, *152*, 106405. [CrossRef]
3. Al Mudawi, N.; Alazeb, A. A model for predicting cervical cancer using machine learning algorithms. *Sensors* **2022**, *22*, 4132. [CrossRef] [PubMed]
4. Biswas, N.; Uddin, K.M.M.; Rikta, S.T.; Dey, S.K. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. *Healthc. Anal.* **2022**, *2*, 100116. [CrossRef]
5. Sabu, K.; Ramnath, M.; Choudhary, A.; Raj, G.; Prakash Agrawal, A. A comparison of traditional and ensemble machine learning approaches for parkinson's disease classification. In *Machine Intelligence and Data Science Applications: Proceedings of MIDAS 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 25–33.
6. Ghimire, S.; Nguyen-Huy, T.; Deo, R.C.; Casillas-Perez, D.; Salcedo-Sanz, S. Efficient daily solar radiation prediction with deep learning 4-phase convolutional neural network, dual stage stacked regression and support vector machine CNN-REGST hybrid model. *Sustain. Mater. Technol.* **2022**, *32*, e00429. [CrossRef]
7. Badjie, B.; Deniz Ülker, E. A Deep Transfer Learning Based Architecture for Brain Tumor Classification Using MR Images. *Inf. Technol. Control* **2022**, *51*, 332–344. [CrossRef]
8. Garzón, A.; Kapelan, Z.; Langeveld, J.; Taormina, R. Machine learning-based surrogate modeling for urban water networks: Review and future research directions. *Water Resour. Res.* **2022**, *58*, e2021WR031808. [CrossRef]
9. Mani, R.C.; Kamalakannan, J. The comparative study of CNN models for breast histopathological image classification. In Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 23–25 January 2023; pp. 1–5.
10. Tummala, S.; Kim, J.; Kadry, S. BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of swin transformers. *Mathematics* **2022**, *10*, 4109. [CrossRef]
11. Liu, Z.; Lv, Q.; Yang, Z.; Li, Y.; Lee, C.H.; Shen, L. Recent progress in transformer-based medical image analysis. *Comput. Biol. Med.* **2023**, *164*, 107268. [CrossRef] [PubMed]
12. Takahashi, S.; Sakaguchi, Y.; Kouno, N.; Takasawa, K.; Ishizu, K.; Akagi, Y.; Aoyama, R.; Teraya, N.; Bolatkan, A.; Shinkai, N.; et al. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *J. Med. Syst.* **2024**, *48*, 84. [CrossRef]
13. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artif. Intell. Rev.* **2023**, *56*, 3005–3054.
14. Casini, L.; Marchetti, N.; Montanucci, A.; Orrù, V.; Roccetti, M. A human–AI collaboration workflow for archaeological sites detection. *Sci. Rep.* **2023**, *13*, 8699. [CrossRef]
15. Asiri, A.A.; Shaf, A.; Ali, T.; Shakeel, U.; Irfan, M.; Mehdar, K.M.; Halawani, H.T.; Alghamdi, A.H.; Alshamrani, A.F.A.; Alqhtani, S.M. Exploring the power of deep learning: Fine-tuned vision transformer for accurate and efficient brain tumor detection in MRI scans. *Diagnostics* **2023**, *13*, 2094. [CrossRef]
16. Zulfiqar, F.; Bajwa, U.I.; Mehmood, Y. Multi-class classification of brain tumor types from MR images using EfficientNets. *Biomed. Signal Process. Control* **2023**, *84*, 104777. [CrossRef]
17. Hossain, S.; Chakrabarty, A.; Gadekallu, T.R.; Alazab, M.; Piran, M.J. Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 1261–1272. [CrossRef]
18. Tummala, S.; Kadry, S.; Bukhari, S.A.C.; Rauf, H.T. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Curr. Oncol.* **2022**, *29*, 7498–7511. [CrossRef] [PubMed]
19. Jiang, Y.; Zhang, Y.; Lin, X.; Dong, J.; Cheng, T.; Liang, J. SwinBTS: A method for 3D multimodal brain tumor segmentation using swin transformer. *Brain Sci.* **2022**, *12*, 797. [CrossRef] [PubMed]

20. ZainEldin, H.; Gamel, S.A.; El-Kenawy, E.S.M.; Alharbi, A.H.; Khafaga, D.S.; Ibrahim, A.; Talaat, F.M. Brain tumor detection and classification using deep learning and sine-cosine fitness grey wolf optimization. *Bioengineering* **2022**, *10*, 18. [CrossRef] [PubMed]
21. Zhang, Y.; Ngo, H.C.; Zhang, Y.; Yusof, N.F.A.; Wang, X. Imaging Segmentation of Brain Tumors Based on the Modified U-net Method. *Inf. Technol. Control* **2024**, *53*, 1074–1087. [CrossRef]
22. Odusami, M.; Maskeliūnas, R.; Damaševičius, R.; Misra, S. Explainable Deep-Learning-Based Diagnosis of Alzheimer's Disease Using Multimodal Input Fusion of PET and MRI Images. *J. Med. Biol. Eng.* **2023**, *43*, 291–302. [CrossRef]
23. Tehsin, S.; Nasir, I.M.; Damaševičius, R.; Maskeliūnas, R. DaSAM: Disease and Spatial Attention Module-Based Explainable Model for Brain Tumor Detection. *Big Data Cogn. Comput.* **2024**, *8*, 97. [CrossRef]
24. Ullah, M.S.; Khan, M.A.; Albarakati, H.M.; Damaševičius, R.; Alsenan, S. Multimodal brain tumor segmentation and classification from MRI scans based on optimized DeepLabV3+ and interpreted networks information fusion empowered with explainable AI. *Comput. Biol. Med.* **2024**, *182*, 109183. [CrossRef] [PubMed]
25. Nickparvar, M. Brain Tumor MRI Dataset. *KAGGLE* **2021**. [CrossRef]
26. Cheng, J. Brain magnetic resonance imaging tumor dataset. *Figshare MRI Dataset Version* **2017**, *5*. [CrossRef]
27. Rosebrock, A. Finding Extreme Points in Contours with Open CV. 2016. Available online: https://pyimagesearch.com/2016/04/11/finding-extreme-points-in-contours-with-opencv/ (accessed on 24 October 2024)
28. Nitesh, V.C. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321.
29. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
30. Vaswani, A. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
31. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
32. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
33. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
34. Redmon, J. Darknet: Open Source Neural Networks in C. 2013–2016. Available online: http://pjreddie.com/darknet/ (accessed on 24 October 2024).
35. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
36. Bahdanau, D. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
37. Ba, J.L. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Dosovitskiy, A. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
40. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
41. Akhter, A.; Acharjee, U.K.; Talukder, M.A.; Islam, M.M.; Uddin, M.A. A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Nat. Lang. Process. J.* **2023**, *4*, 100027. [CrossRef]
42. Ahmad, W.; Ali, H.; Shah, Z.; Azmat, S. A new generative adversarial network for medical images super resolution. *Sci. Rep.* **2022**, *12*, 9533. [CrossRef] [PubMed]