Open Access

OBM Neurobiology

LIDSEN Publishing Inc.

Original Research

Modern Vision Transformer Models for Accurate and Explainable Brain Tumor MRI Classification

Ishak Pacal [1, 2, 3, *]

1. Department of Computer Engineering, Faculty of Engineering, Igdir University, 76000 Igdir, Turkey; E-Mail: ishakpacal@igdir.edu.tr; ORCID: 0000-0001-6670-2169
2. Department of Electronics and Information Technologies, Faculty of Architecture and Engineering, Nakhchivan State University, Nakhchivan, AZ 7012, Azerbaijan
3. Department of Computer Engineering, Faculty of Engineering and Architecture, Fenerbahce University, Istanbul, Turkey

* **Correspondence:** Ishak Pacal; E-Mail: ishakpacal@igdir.edu.tr; ORCID: 0000-0001-6670-2169

Academic Editor: Fady Alnajjar

OBM Neurobiology
2026, volume 10, issue 3
doi:10.21926/obm.neurobiol.2603352

Received: June 06, 2026
Accepted: August 18, 2026
Published: September 09, 2026

Abstract

Accurate brain tumor classification from magnetic resonance imaging (MRI) remains difficult because glioma, meningioma, and pituitary tumors may exhibit overlapping enhancement, morphology, and slice-dependent appearance. This study presents a controlled comparison of Vision Transformer, Data-efficient Image Transformer (DeiT), Swin Transformer, BEiT, and EVA-02 under a common image-level protocol. The public Epic and CSCR Hospital Dataset record describes 12,064 preprocessed T1-weighted contrast-enhanced MRI images and was evaluated according to its published partitioning scheme. The original test set was reserved for final evaluation, and a validation subset was drawn solely from the original training set. BEiT-Base produced the highest observed accuracy (0.9921; Wilson 95% confidence interval, 0.9877-0.9950), a macro F1-score of 0.9924, and a macro area under the receiver operating characteristic curve of 0.9995, at a profiled cost of 33.70 giga floating-point operations per image, equal to that of Vision Transformer-Base and DeiT-Base. Accuracy intervals overlapped across models, and the differences were interpreted descriptively. Gradient-weighted Class Activation Mapping showed activation near the visually apparent lesion in many cases, but

peripheral and non-lesion responses were also observed. Because patient and examination identifiers are unavailable, the findings represent image-level benchmark performance and do not establish patient-level generalization.

Keywords
Brain tumor classification; magnetic resonance imaging; controlled benchmark; vision transformer; image-level evaluation; explainable artificial intelligence; Grad-CAM

1. Introduction

Brain tumors are among the most consequential neurological diseases because even a relatively small intracranial lesion can disrupt cognition, movement, language, vision, endocrine function, or intracranial pressure. Their clinical effects depend on anatomical location, growth pattern, molecular characteristics, and interaction with densely connected neural networks [1]. Gliomas are particularly difficult to manage because of their infiltrative growth, biological heterogeneity, and treatment resistance, and current classification integrates molecular findings with histological and imaging features [2, 3].

The clinical burden remains substantial. The American Cancer Society projected 2,114,850 new cancer cases and 626,140 cancer deaths in the United States in 2026. Brain and other nervous system cancers were expected to account for 24,740 new cases and 18,350 deaths, reflecting a high mortality burden relative to their incidence [4].

Magnetic resonance imaging (MRI) is central to brain tumor assessment because it provides high soft-tissue contrast and depicts lesion location, margins, enhancement, internal heterogeneity, mass effect, and relationships with adjacent structures. Automated classification remains challenging because glioma, meningioma, and pituitary tumors may share local visual characteristics. At the same time, scanner variation, acquisition differences, anatomical plane, and the restricted context of a single two-dimensional slice can alter their appearance [5, 6]. Clinical interpretation also draws on volumetric information and complementary sequences, including T2-weighted, fluid-attenuated inversion recovery (FLAIR), diffusion-weighted imaging, and pre- and post-contrast T1-weighted images. Classification from isolated two-dimensional slices from this provider-described T1-weighted contrast-enhanced benchmark therefore represents a limited image-recognition task rather than the complete diagnostic process.

Recent deep learning studies have reported strong performance in breast lesion classification and ischemic stroke imaging [7-9]. In parallel, endoscopic super-resolution methods have emphasized edge preservation and computational efficiency under low-resolution conditions [10, 11]. Although these applications address different clinical tasks, they illustrate the importance of modality-specific model design, anatomical preservation, and rigorous validation in medical imaging.

Convolutional neural networks (CNNs) remain effective medical-image classifiers because local receptive fields encode edges, texture, and shape efficiently. Transformer architectures complement this local representation by modeling relationships among spatially separated image regions through self-attention [12, 13]. Their ability to capture broader context provides a reason for systematic evaluation in brain MRI, but it does not imply an inherent advantage over CNNs or

lightweight hybrid models. Recent studies have reported competitive results from attention-guided CNNs, compact architectures, and CNN-transformer designs [14, 15]. Direct comparison across studies remains unreliable because datasets, partitioning strategies, augmentation, class balancing, validation procedures, and test-time processing differ substantially.

The present study was designed as a controlled benchmark rather than a new diagnostic framework. It examines whether five established transformer families differ when data organization, preprocessing, augmentation, optimization, checkpoint selection, and final evaluation are kept constant. The analysis focuses on three questions: whether the models differ meaningfully in predictive performance, whether any gain is accompanied by lower computational demand, and which errors or attention failures remain visible at the image level.

Experiments were conducted on the public Epic and CSCR Hospital Dataset, whose Mendeley Data record describes 12,064 preprocessed T1-weighted contrast-enhanced MRI images assigned to glioma, meningioma, pituitary tumor, and no-tumor classes [16]. The study retained the published image-level split: the original test set was reserved for final inference, and only the original training set was divided to obtain a validation subset. The dataset is distributed without patient or examination identifiers, so the subsets cannot be confirmed to separate patients or examinations and may contain related slices. The high scores are therefore interpreted as benchmark-level, rather than patient-level, performance.

The analysis compares five transformer families under one downstream protocol and reports classification performance, computational cost, confidence intervals, error patterns, and qualitative activation maps. This design provides a consistent image-level reference for the public benchmark while keeping the conclusions' scope aligned with the available data.

2. Related Work

2.1 CNN, Attention, Hybrid, and Efficiency-Oriented Approaches

CNN-based brain tumor classifiers have been extended with multiscale kernels, attention, feature fusion, classical classifiers, and lightweight operators. Khan et al. introduced ShallowMRI, which combines contour-oriented preprocessing, texture descriptors, and an attention-guided lightweight CNN, reporting 98.24% accuracy on a multiclass Kaggle setting [14]. Binish et al. integrated channel-spatial attention with separable multi-resolution kernels [17], while Hasan et al. coupled strip-style pooling attention with established CNN backbones and used Grad-CAM variants for visual interpretation [18]. Other studies have explored synthetic feature fusion, hypergraph attention, three-branch convolutional designs, multiscale dense extraction, and CNN-support vector machine combinations [19-23].

Recent compact convolutional and hybrid models also show why a transformer-only benchmark cannot establish transformer superiority. HayabusaNet uses depthwise separable convolution, parallel multiscale branches, feature fusion, and hybrid attention; it reported 98.86% accuracy with 1.319 million parameters on a 7,023-image four-class dataset and evaluated explanations with Grad-CAM and Grad-CAM++ [24]. TumorNet combines a compact convolutional front end with MobileViT-XXS and reported 99.85% accuracy with 1.26 million trainable parameters, together with tests on two additional public datasets [25]. DiSCNet reported 0.9922 accuracy with 2.78 million parameters using directional split convolution [26]. These studies show that compact CNN and hybrid designs can match or exceed the accuracy of much larger transformers on their respective

datasets. They also confirm that parameter efficiency and external testing are separate dimensions from peak within-dataset accuracy.

Table 1 summarizes results obtained with different datasets and evaluation procedures. These values provide context for accuracy and model size, but they do not permit a direct ranking of architecture families.

Table 1 Comparative synthesis of representative recent brain tumor MRI classifiers. Values originate from different datasets and evaluation protocols and must not be interpreted as a direct ranking. Values that were unavailable in the cited source are listed as not reported.

Study	Model family	Dataset and evaluation	Accuracy	Parameters (million)	Explainability or validation feature
Khan et al. [14]	Lightweight attention CNN	Multiclass Kaggle setting	98.24%	Not reported	Attention-guided feature selection
Prayogo et al. [24]	Multiscale attention CNN	Kaggle, 7,023 images	98.86%	1.319	Grad-CAM and Grad-CAM++; Br35H analysis
Wang et al. [25]	CNN-MobileViT hybrid	Primary plus two public test datasets	99.85%	1.26	External-dataset tests; qualitative explanations
Ganie and Pacal [26]	Compact directional CNN	Merged four-class MRI dataset	99.22%	2.78	Efficiency-oriented evaluation
Present benchmark	Five transformer families	Epic and CSCR predefined image-level test	99.21%[a]	85.77	Wilson interval; qualitative Grad-CAM

[a]Highest observed accuracy point estimate among the five transformer models; no claim of statistical superiority.

2.2 Transformer-Based, Explainable, and Broader Medical-Image Approaches

The Vision Transformer introduced global token-based self-attention for image recognition [12]. Data-efficient Image Transformer (DeiT) improved training efficiency for data-limited settings [27]; Swin Transformer introduced hierarchical shifted-window attention [13]; BEiT used masked image modeling for visual pretraining [28]; and EVA-02 extended large-scale visual representation learning [29]. In brain tumor MRI, transformer-based and hybrid studies have reported competitive performance, but their conclusions remain sensitive to data organization and model-development choices [15, 30].

Explainability and deployment have become parallel evaluation goals. Multi-stage attention using medical-foundation features has been proposed to reduce computational demand [31], and smartphone-oriented studies have examined on-device inference [32]. Fuzzy decision trees offer an explainable-by-design alternative to black-box networks [33], while active learning has been investigated to reduce annotation requirements [34]. Grad-CAM remains widely used because it

maps class-sensitive activations back to image space. Still, it is a post-hoc visualization rather than proof that a model has learned a clinically valid causal mechanism [35].

Accordingly, the study is restricted to a within-transformer comparison under a common image-level protocol. Performance, computational cost, classification errors, and qualitative activation patterns are evaluated without extrapolating the findings to patient-level or clinical performance.

2.3 Ethics Approval and Consent to Participate

This study used the publicly available Epic and CSCR Hospital Dataset distributed through Mendeley Data. The dataset consists of de-identified brain MRI images and does not include patient names, hospital registration numbers, demographic identifiers, or directly identifiable personal information. Since this work is a secondary computational analysis of a publicly available de-identified dataset, additional institutional ethics approval was not required. Consent to participate was not applicable because the study did not involve direct patient recruitment, prospective data collection, clinical intervention, or contact with human participants.

3. Materials and Methods

3.1 Dataset and Classification Task

The experiments were conducted on the Epic and CSCR Hospital Dataset, a public four-class brain MRI benchmark distributed through Mendeley Data [16]. The Mendeley record lists the collection as the Brain Tumor MRI Dataset (Glioma, Meningioma, Pituitary, No Tumor) and describes all four class folders ccontain preprocessed T1-weighted contrast-enhanced MRI images. The distributed files are JPEG/PNG images rather than DICOM acquisitions and do not provide image-level sequence tags or acquisition headers. Sequence identity was therefore not independently reassigned from the image files; the sequence terminology used in this study follows the description supplied by the dataset provider.

The benchmark provides image-level class labels and a predefined train-test split. Patient and examination identifiers, scanner parameters, acquisition protocols, demographic variables, and center information are unavailable. The published split was therefore retained without attempting to infer patient groups from image names. The evaluation is image-level, and independence at the patient or examination level cannot be verified.

The dataset contains 12,064 MRI images. The original test set remained unchanged, while the original training set was divided into training and validation subsets. The class distribution is reported in Table 2, and representative images are shown in Figure 1.

Table 2 Class-wise distribution of the Epic and CSCR Hospital Dataset used in this image-level benchmark.

Class	Training	Validation	Benchmark test	Total
Glioma	2,414	604	755	3,773
Meningioma	1,746	437	546	2,729
Pituitary Tumor	2,003	501	626	3,130
No Tumor	1,556	389	487	2,432
Total	7,719	1,931	2,414	12,064

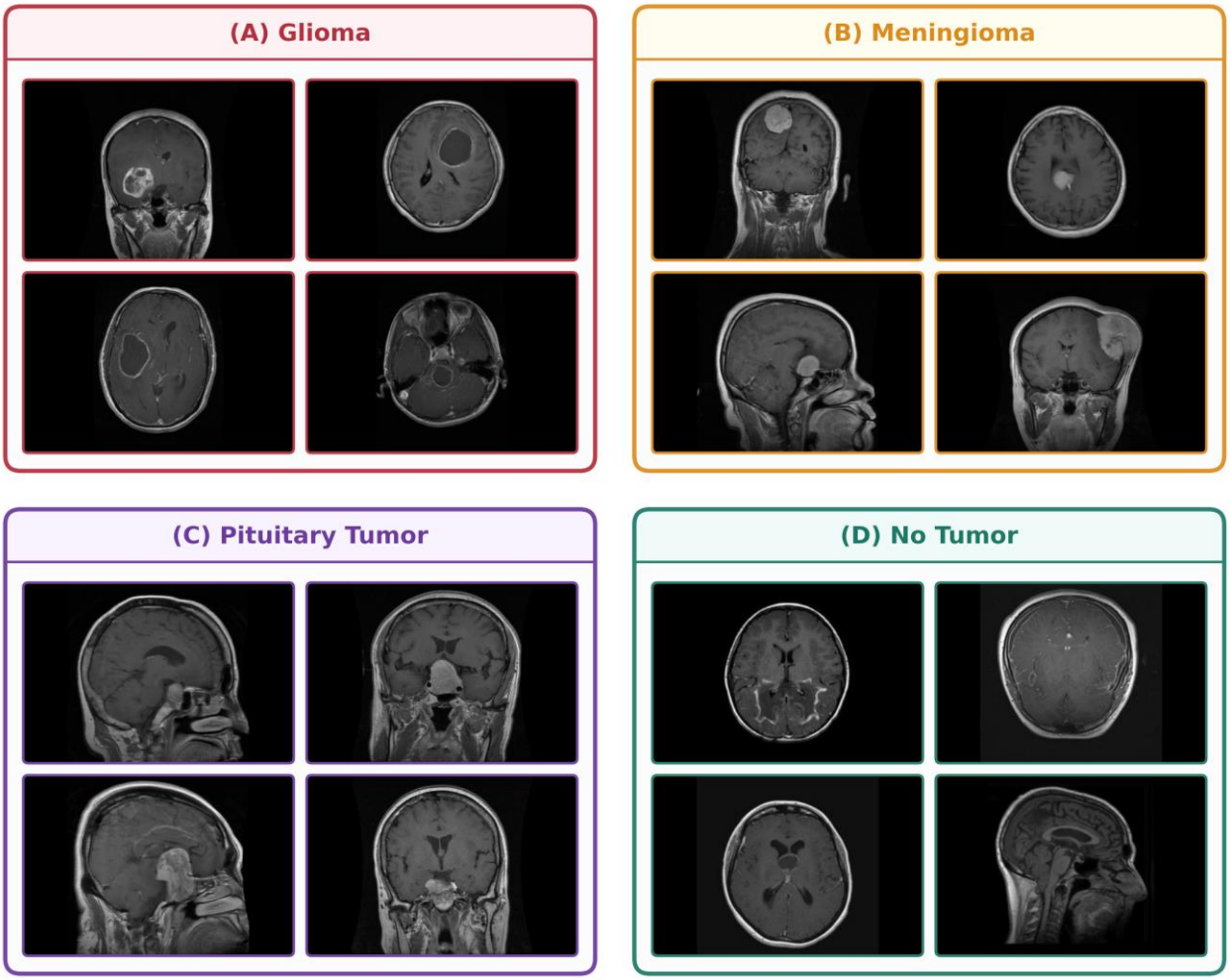

Figure 1 Representative MRI images from the Epic and CSCR Hospital Dataset across glioma, meningioma, pituitary tumor, and no-tumor classes. The dataset provider describes the released images in all four class folders as preprocessed T1-weighted contrast-enhanced MRI. The examples illustrate variation in anatomical plane, lesion location, morphology, enhancement pattern, and image contrast.

The distribution indicates mild class imbalance. Glioma and pituitary tumor samples are more frequent than no-tumor images, whereas meningioma samples fall into an intermediate group. Accuracy was therefore interpreted together with macro-averaged metrics.

3.2 Data Partitioning and Preprocessing

The original test set was reserved for final evaluation and was excluded from preprocessing decisions, hyperparameter adjustment, validation, checkpoint selection, and model comparison. Only the original training set was divided into training and validation subsets. This procedure follows the published benchmark organization; it is not a patient-based split because patient and examination identifiers are unavailable.

All MRI slices were converted into a fixed input format compatible with the evaluated pretrained backbones. Each image was resized to 224 × 224, replicated to three channels when required, and

normalized according to the preprocessing configuration of the corresponding backbone. For a resized image I^r, intensity normalization was defined as

$$I^n_{u,v,c} = \frac{I^r_{u,v,c} - \mu_c}{\sigma_c}, \quad (1)$$

where $I^n_{u,v,c}$ denotes the normalized intensity at pixel location (u, v) and channel c, while μ_c and σ_c are the channel-wise mean and standard deviation.

No skull stripping, tumor segmentation, lesion cropping, handcrafted texture extraction, or radiomic feature engineering was applied. The benchmark therefore examines end-to-end image classification by pretrained transformer families rather than the effect of an additional region-of-interest or radiomics pipeline.

Figure 2 summarizes the experimental workflow. Each model uses the same image partitions and evaluation procedure while retaining its native tokenization, representation, pooling, and normalization operations. Final test-set evaluation includes accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUROC), confusion matrices, parameter count, and giga floating-point operations (GFLOPs).

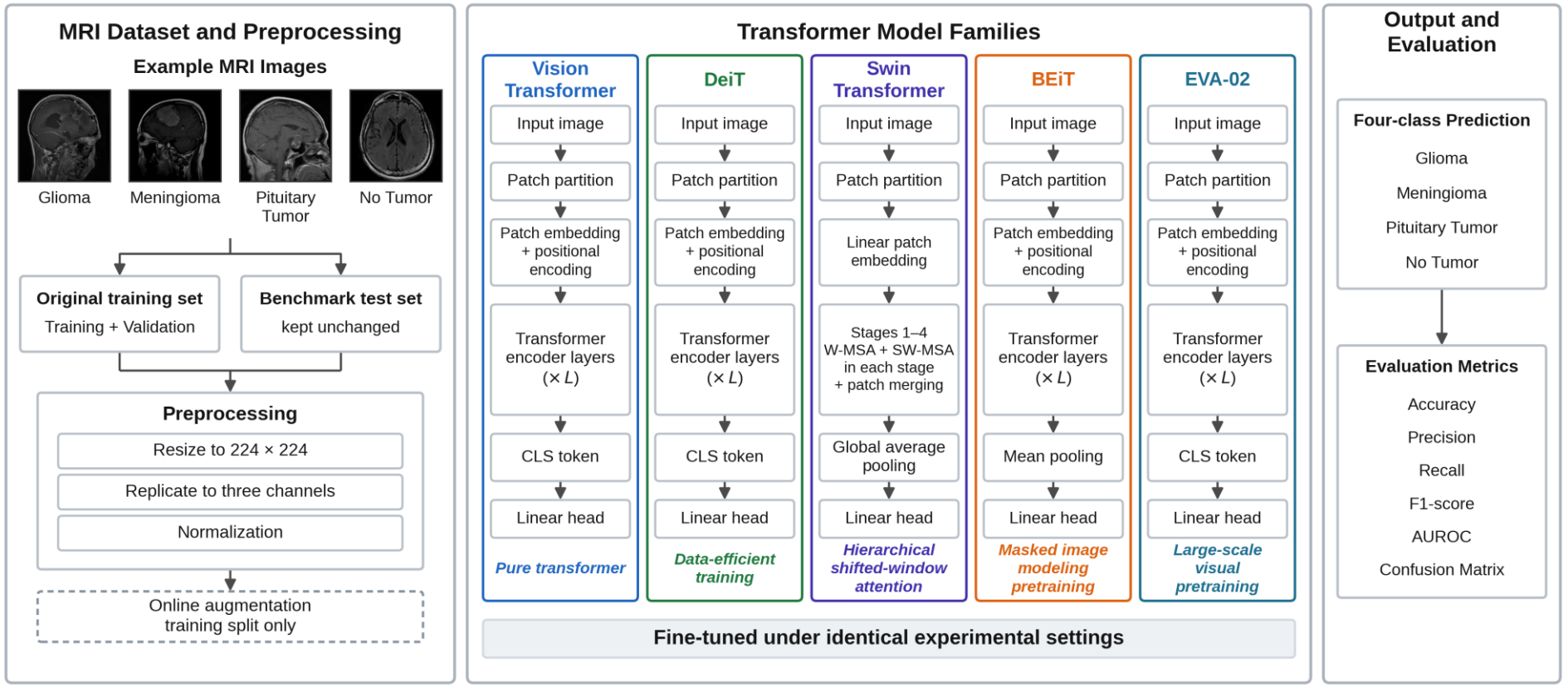

Figure 2 Overview of the controlled comparison. Left: four example slices from the public Epic and CSCR Hospital Dataset, one per class. The published partition was retained, so the benchmark test set was held out before any preprocessing decision was made, and only the original training set was divided into training and validation subsets. Resizing, three-channel replication and normalization were applied to all images, whereas augmentation was applied online during training to the training split only. Middle: the five transformer families, each fine-tuned under identical data partitions, preprocessing, optimization, schedule and checkpoint selection. Right: the four-class prediction task and the measures computed on the predefined benchmark test set.

3.3 Transformer Families and Classification Formulation

Five pretrained model families were evaluated: Vision Transformer [12], Data-efficient Image Transformer (DeiT) [27], Swin Transformer [13], BEiT [28], and EVA-02 [29]. They differ in their pretraining objectives, global or windowed attention, and hierarchical representation, but all transform an image into a sequence or hierarchy of visual tokens. Given an MRI image $x \in R^{H \times W \times C}$, a canonical patch-based representation can be written as

$$z_0 = \left[x_{cls}; x_p^1 E; x_p^2 E; \ldots; x_p^n E\right] + E_{pos}, \qquad N = \frac{HW}{p^2}, \tag{2}$$

where x_{cls} is a learnable class token, x_p^j is the j-th image patch, E is the patch projection matrix, E_{pos} is the positional embedding, P denotes patch size, and N is the number of patches.

For a token matrix Z, scaled dot-product attention is computed from query, key, and value projections:

$$Q = ZW_Q, \qquad K = ZW_K, \qquad V = ZW_V, \qquad Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{3}$$

A transformer encoder block updates the token sequence through multi-head self-attention (MSA), layer normalization, a multilayer perceptron, and residual connections. Let N denote layer normalization and M denote the multilayer perceptron:

$$z_l' = MSA\big(N(z_{l-1})\big) + z_{l-1}, \qquad z_l = M\big(N(z_l')\big) + z_l'. \tag{4}$$

For each backbone, the native final pooling or class-token selection and pre-classifier normalization were retained. The original final classifier was replaced by one newly initialized affine layer with bias, mapping the backbone-specific embedding $h_i \in R^{d_m}$ to four logits: $s_i = W_m h_i + b_m$, where $W_m \in R^{4 \times d_m}$ and $b_m \in R^4$. No additional hidden layers or task-specific attention modules were added. The predicted probability for class c is

$$p(y_i = c|x_i) = \frac{exp(s_{i,c})}{\sum_{k=1}^{C} exp(s_{i,k})}, \qquad C = 4. \tag{5}$$

Vision Transformer provides the canonical global self-attention baseline. DeiT was included because its training formulation was designed to improve data efficiency. Swin Transformer computes attention within shifted local windows and forms hierarchical features. BEiT uses masked image modeling during pretraining, whereas EVA-02 uses large-scale visual representation learning. These design differences motivate a controlled within-family comparison, but, by themselves, they do not establish suitability relative to CNN or hybrid architectures.

3.4 Transfer Learning and Data Augmentation

All models were initialized with pretrained weights and fine-tuned for the four image-level classes. Let ϕ_{θ_b} denote the pretrained backbone and g_{θ_c} the newly initialized linear classification layer. The complete model is

$$f_\theta(x_i) = g_{\theta_c}(\phi_{\theta_b}(x_i)). \tag{6}$$

Label smoothing was used to reduce overconfident optimization. For ε = 0.1, the target distribution was

$$\tilde{y}_{i,c} = (1-\varepsilon)y_{i,c} + \frac{\varepsilon}{C}, \tag{7}$$

and the label-smoothed cross-entropy objective was

$$\boldsymbol{L}_{classification} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}\tilde{y}_{i,c}\, log\big(p_{i,c}\big), \tag{8}$$

where M is the mini-batch size and C = 4.

Augmentation was applied only to training images. The implemented pipeline used random resized cropping to 224 × 224 with a scale range of 0.08-1.00 and an aspect-ratio range of 0.75-1.33, horizontal flipping with probability 0.5, color jitter with magnitude 0.4, and random interpolation during resizing. Vertical flipping, random erasing, repeated augmentation, augmentation splits, Gaussian blur, grayscale augmentation, and test-time augmentation were disabled. Validation and test images were processed deterministically and evaluated once.

The lower crop bound of 0.08 originated from a standard natural-image fine-tuning recipe and was not optimized for MRI anatomy. In an isolated two-dimensional slice, such a crop can remove part or all of a lesion or alter its anatomical context. Tumor masks were not provided, so lesion-retention frequency could not be quantified. The augmentation settings are reported as implemented; however, this policy was not optimized or validated for anatomical preservation in brain MRI.

3.5 Performance Evaluation

Final performance was measured on the unchanged predefined test set. Let TP_c, TN_c, FP_c, and FN_c denote true positives, true negatives, false positives, and false negatives for class c under a one-versus-rest formulation. Accuracy and macro-averaged metrics were computed as

$$\begin{aligned}
Accuracy &= \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C}(TP_c + FN_c)}, \\
Precision_{macro} &= \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c}, \\
Recall_{macro} &= \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c}, \\
F1_{macro} &= \frac{1}{C}\sum_{c=1}^{C}\frac{2\cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c}, \\
AUROC_{macro} &= \frac{1}{C}\sum_{c=1}^{C} AUROC_c\,.
\end{aligned} \tag{9}$$

Macro-averaging assigns equal weight to each class. Model complexity was assessed using trainable parameter count and computational profiling with THOP. For each architecture, THOP was applied to a single input of size 1 × 3 × 224 × 224 and returned multiply-accumulate operations (MACs). The reported GFLOPs were obtained using the convention 1 MAC = 2 FLOPs and dividing the resulting operation count by 10^9. The same profiling procedure and input dimensions were used for all five models. Because timm's BEiT applies the qkv projection functionally, an explicit operation handler was registered for that module to ensure the same layer types were counted across all models.

Wilson 95% confidence intervals were computed for the accuracy of all five models and for class-wise recall of BEiT-Base. The predefined test set remained fixed, and validation data were obtained only from the released training portion. The experiment used one random seed. Repeated runs and paired prediction tests were not performed; therefore, the comparison is based on point estimates and confidence intervals rather than claims of statistical superiority.

3.6 Grad-CAM Analysis

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied after final evaluation to the model with the highest observed accuracy point estimate [35]. The analysis included eight correctly classified examples across the four classes and four author-selected examples with predominantly peripheral or non-lesion activation. For tumor classes, the maps were examined descriptively in relation to the visually apparent abnormality, whereas no focal lesion target was assumed for the no-tumor class. Eight classification errors are displayed separately in the prediction panel. Heatmaps were not used for training, hyperparameter selection, or checkpoint selection.

For class *c*, Grad-CAM was computed from a selected activation tensor A^k and class score s^c as

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial s^c}{\partial A_{ij}^k}, \qquad L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right), \tag{10}$$

where α_k^c is the importance weight of feature map *k*, *Z* is the number of spatial locations, and $L_{Grad-CAM}^c$ is the localization map. The maps were resized to image resolution and overlaid on the MRI slices.

The dataset contains no tumor masks, bounding boxes, or expert localization ratings, and independent radiologist review was not performed for the Grad-CAM assessment. Quantitative overlap measures and expert agreement could therefore not be assessed. The heatmaps are interpreted descriptively and are not used as evidence of localization accuracy.

4. Results and Discussion

4.1 Experimental Setup

All experiments were implemented in Python 3.14 using PyTorch 2.12 with NVIDIA CUDA 13.0 support. The workstation contained an NVIDIA RTX 4090 graphics processing unit (GPU), an Intel Core i9-14900K processor, and 64 GB of DDR5 random-access memory (RAM). The same software environment and hardware configuration were used for all evaluated model families.

All images were resized to224 × 224, and the number of output classes was set to four. The evaluated models were initialized with pretrained weights and adapted to the brain MRI classification task by replacing the original classifier with a four-class task-specific head. The test set was not used during training, validation, checkpoint selection, or any model-development stage.

4.2 Training Details

All transformer families were trained under the same optimization protocol. The batch size was 16 and the maximum duration was 300 epochs. Early stopping was applied with a patience of 50 epochs; training was terminated when the monitored validation performance did not improve for 50 consecutive epochs. Stochastic gradient descent was used with momentum 0.9 and weight decay 2.0×10^{-5}. A cosine learning-rate schedule used 5 warm-up epochs, an initial warm-up learning rate of 1.0×10^{-5}, and a base learning-rate setting of 0.1. The learning rate configuration was applied identically to all pretrained transformer backbones, ensuring architecture comparisons were made under a common optimization schedule.

Mixed-precision training, gradient accumulation, gradient checkpointing, model exponential moving average, mixup, CutMix, random erasing, and test-time augmentation were disabled. Label smoothing was 0.1 and the random seed was fixed at 42.

Training and validation accuracy and loss trajectories were inspected throughout optimization to assess convergence and possible overfitting; early stopping provided an additional safeguard against prolonged training after validation performance had plateaued.

4.3 Quantitative Comparison of Transformer Families

Table 3 reports performance on the predefined test set. Precision, recall, F1-score, and AUROC are macro-averaged. All models achieved high image-level scores, with accuracies above 0.987 and AUROCs above 0.998.

Table 3 Image-level test performance of the five transformer families. Accuracy is accompanied by Wilson 95% confidence intervals. GFLOPs were derived from THOP-reported MACs for a 1 × 3 × 224 × 224 input using the convention 1 MAC = 2 FLOPs. Counting covers linear and convolutional layers.

Model	Accuracy	95% confidence interval	Precision	Recall	F1-score	Parameters (million)	GFLOPs	AUROC
Vision Transformer-Base	0.9888	0.9838-0.9923	0.9883	0.9902	0.9892	85.80	33.70	0.9981
DeiT-Base	0.9872	0.9818-0.9909	0.9859	0.9890	0.9874	85.80	33.70	0.9997
Swin-Base	0.9880	0.9828-0.9916	0.9876	0.9886	0.9881	86.75	30.34	0.9997
BEiT-Base	0.9921	0.9877-0.9950	0.9919	0.9929	0.9924	85.77	33.70	0.9995
EVA-02-Base	0.9905	0.9857-0.9936	0.9898	0.9918	0.9908	85.76	43.89	0.9992

BEiT-Base produced the highest observed accuracy and macro F1 point estimates at the same profiled operation count as Vision Transformer-Base and DeiT-Base, which share its patch-16 base configuration. Swin-Base had the lowest profiled operation count among the five base-sized transformers. EVA-02-Base produced the second-highest accuracy point estimate but had the greatest profiled operation count. DeiT-Base and Swin-Base produced the highest AUROC point estimates. The small numerical differences do not establish a statistically significant ranking.

The accuracy intervals overlap substantially. In particular, the BEiT-Base interval of 0.9877-0.9950 overlaps the intervals of every other model. The interval indicates that the test accuracy estimate is precise within the released image-level subset. Still, it does not address patient overlap, dataset shift, training-run variability, or external generalization. For BEiT-Base, class-wise recall intervals were 0.9724-0.9909 for glioma, 0.9787-0.9961 for meningioma, 0.9922-1.0000 for the no-tumor class, and 0.9884-0.9991 for pituitary tumor.

4.4 Confusion Matrix and Receiver Operating Characteristic Analysis

The class-wise behavior of BEiT-Base, which had the highest observed accuracy point estimate, is presented in Figure 3. The model correctly identified all 487 no-tumor test images, corresponding to a class-wise recall of 1.0000. However, two glioma images were assigned to the no-tumor category, giving the no-tumor class a precision of 487/489 (0.9959) rather than error-free separation. Pituitary tumor showed a class-wise recall of 624/626 (0.9968), glioma 743/755 (0.9841), and meningioma 541/546 (0.9908).

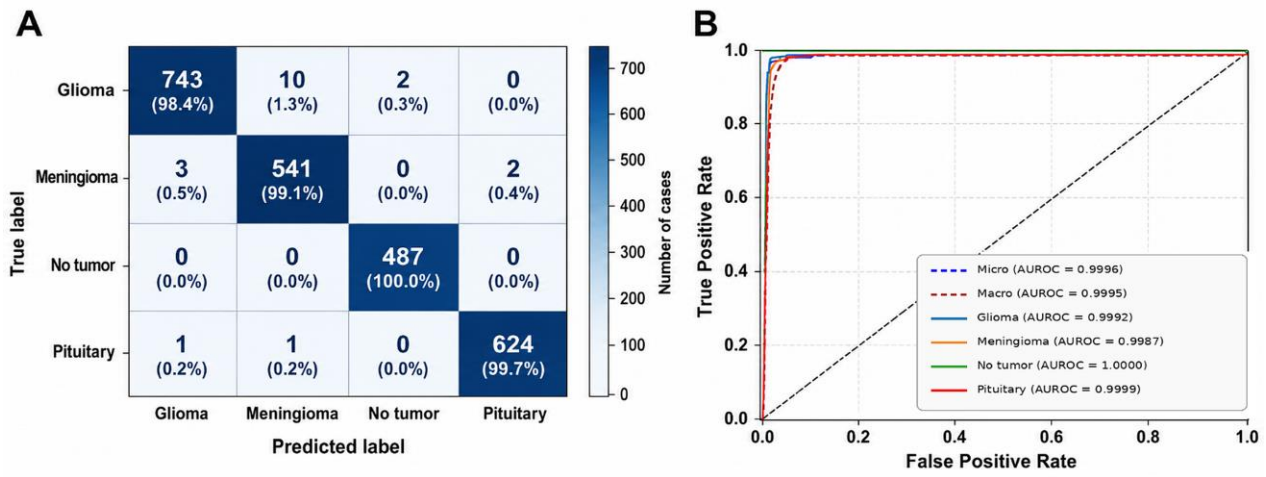

Figure 3 Confusion matrix and receiver operating characteristic analysis of BEiT-Base on the predefined test set. (A) Confusion matrix showing absolute counts and row-wise percentages for glioma, meningioma, no-tumor, and pituitary tumor classes. (B) One-versus-rest curves with class-wise, micro-averaged, and macro-averaged area under the receiver operating characteristic curve (AUROC) values.

Most errors occurred between tumor classes. Glioma was often misdiagnosed as meningioma, while a small number of meningioma samples were misdiagnosed as glioma or pituitary tumor. These errors are plausible in a slice-level setting because glioma and meningioma may share mass-like appearance, contrast enhancement, and boundary characteristics in individual two-dimensional slices.

The receiver operating characteristic curves in Figure 3 show high ranking separability across all four classes. The macro-averaged AUROC was 0.9995, while the micro-averaged AUROC was 0.9996. Class-wise AUROC values were 0.9992 for glioma, 0.9987 for meningioma, 1.0000 for the no-tumor class, and 0.9999 for pituitary tumor. The no-tumor AUROC of 1.0000 reflects ranking performance and should not be read as error-free classification: the class had complete recall, but two glioma images were falsely predicted as no tumor.

4.5 Prediction-Level Visual Assessment

Representative correct and incorrect predictions are shown in Figure 4. The panel contains correctly classified samples from all four classes and eight misclassified images. Softmax scores are reported as the model's outputs. They should be read against the training objective: label smoothing with ε = 0.1 over four classes places the target for the correct class at 0.925, so confident correct predictions concentrate near that value rather than approaching 1.0. The scores are therefore not calibrated probabilities, and calibration was not evaluated. One misclassified image in Figure 4 had a higher softmax score than any of the correctly classified examples shown, illustrating that a high score does not by itself indicate a reliable prediction. Glioma and meningioma were recognized across different anatomical planes and lesion appearances. Pituitary tumor examples were identified in sellar or parasellar regions, while no-tumor images were correctly classified despite variation in anatomical plane, slice level, and image contrast.

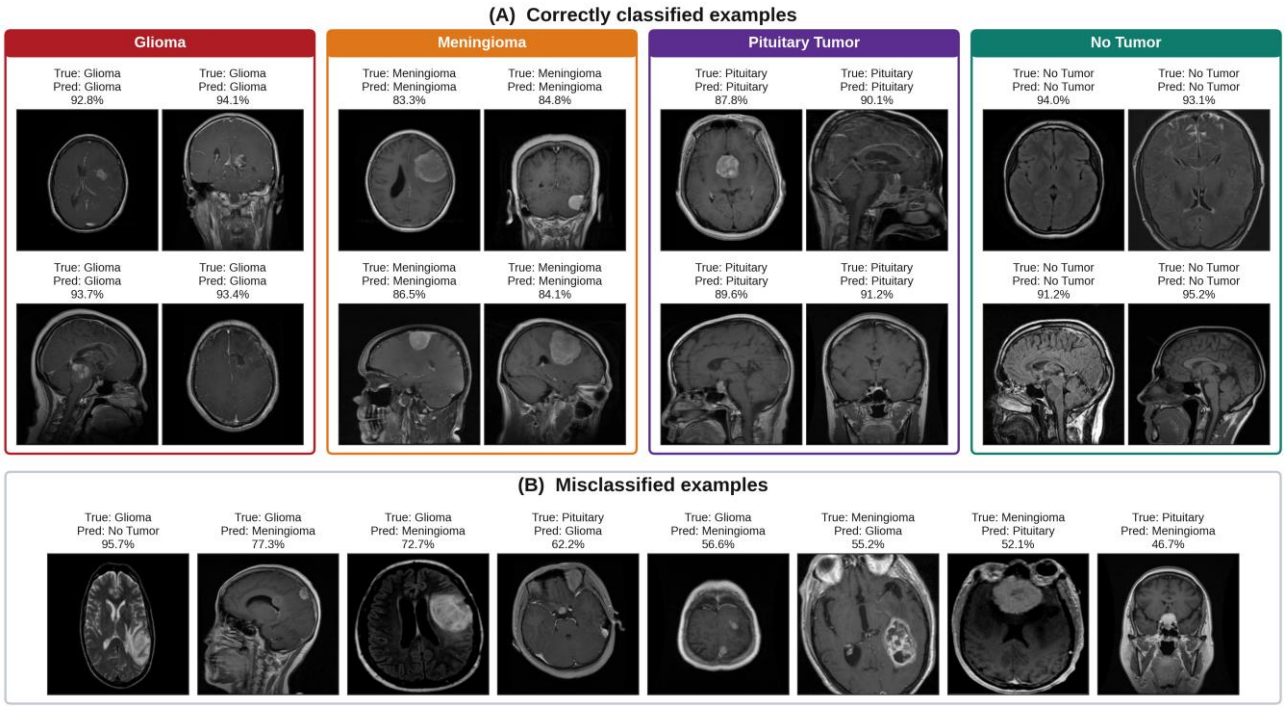

Figure 4 Representative correctly classified and misclassified MRI examples from the test set. (A) Correctly classified samples from glioma, meningioma, pituitary tumor, and no-tumor classes. (B) Misclassified samples showing the true class, predicted class, and confidence score.

The misclassified samples in Figure 4 reveal the remaining failure patterns. Several glioma images were predicted as meningioma or no-tumor, and a limited number of meningioma and pituitary

tumor samples were confused with other tumor categories. The examples show that errors can occur when a single slice provides limited anatomical context or when enhancement and morphology overlap across tumor classes. Because patient history, volumetric context, and additional MRI sequences were unavailable, the visual cause of an error cannot be established from the benchmark image alone.

4.6 Grad-CAM-Based Explainability Analysis

Figure 5 presents eight correctly classified examples across the four classes, as well as four author-selected examples with predominantly peripheral or non-lesion activation. Several glioma, meningioma, and pituitary tumor maps showed activation near the visually apparent enhancing mass, whereas no-tumor maps were more diffuse because no focal lesion target was assumed. The examples describe activation patterns but do not establish localization accuracy.

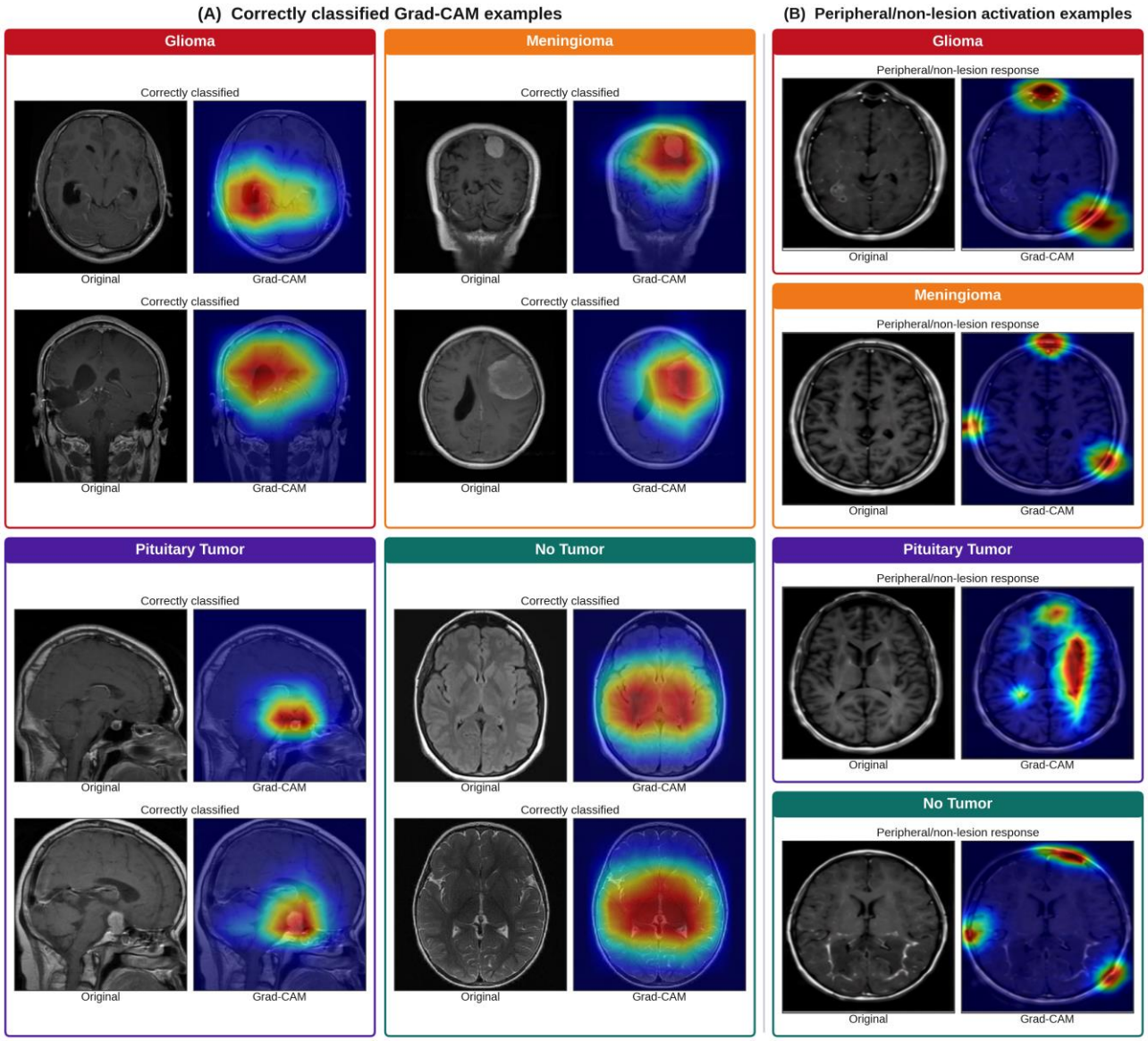

Figure 5 Qualitative Grad-CAM assessment of BEiT-Base. (A) Eight correctly classified examples across the four classes. For tumor classes, the maps are shown in relation to the visually apparent abnormality, whereas no focal lesion target is assumed for the no-

tumor class. (B) Four author-selected examples showing activation concentrated near image boundaries, peripheral regions, or other anatomically implausible areas. Tumor masks and expert localization scores were unavailable; therefore, the figure does not provide quantitative validation of localization.

Some correctly classified images showed activation away from the visually apparent lesion. Such cases indicate that a correct label does not necessarily correspond to a clinically meaningful spatial explanation. Quantitative assessment would require lesion annotations and independent expert review, neither of which was available for this dataset.

4.7 Discussion

All five transformer families achieved high image-level performance under the same experimental conditions, and the numerical differences were small. BEiT-Base produced the highest accuracy and macro F1-score in the present run, but its Wilson interval overlapped those of the other models. In the absence of repeated runs and paired prediction tests, the results support a cautious interpretation: several base-sized transformer families perform similarly on this dataset, rather than one architecture showing a decisive advantage.

Computational demand did not increase in parallel with accuracy. BEiT-Base combined the highest observed accuracy with the same profiled operation count as the other patch-16 base models, whereas EVA-02 required the greatest computation. This finding is limited to the five transformer configurations studied here. Recent compact CNN and hybrid models, including HayabusaNet and TumorNet, report strong results with approximately 1.3 million parameters, compared with approximately 86 million parameters for the present models [24, 25]. Because these studies used different datasets and evaluation procedures, their results provide context rather than a direct comparison. Same-protocol CNN and hybrid experiments are still required before any conclusion can be drawn about the architecture class.

The no-tumor class achieved 100% sensitivity in the test subset, with 487/487 images correctly recognized, but two glioma images were predicted as no-tumor. Its recall was therefore 1.0000, and its precision was 0.9959. Most remaining errors occurred between glioma and meningioma, where isolated slices may provide limited information about lesion origin and extent. The Grad-CAM panels contained both lesion-adjacent and peripheral responses, indicating that strong classification scores can coexist with weak spatial explanations.

The predefined test set was excluded from model selection. Nevertheless, patient and examination identifiers are unavailable, and the distribution of related slices across the published subsets cannot be assessed. This uncertainty should be taken into account when interpreting the high image-level scores.

The augmentation strategy introduces a further source of uncertainty. Random resized cropping with a lower scale of 0.08 can remove part or all of a lesion and alter anatomical context. Because lesion annotations are unavailable, the frequency and influence of such crops could not be measured.

Published transformer, CNN, hybrid, and foundation-model studies differ in dataset composition, partitioning, augmentation, and evaluation. Table 1 should therefore be read as a comparative overview rather than a state-of-the-art ranking. Direct architecture-level conclusions require same-protocol baselines and repeated evaluation.

5. Limitations and Future Work

The main limitation is the absence of patient and examination identifiers. The published image-level split was retained, but independence across patients cannot be verified, and related slices may occur in more than one subset. This uncertainty may contribute to the high scores and limits their interpretation for independent patients.

The experiment used a single random seed and a single final run per model. The predefined test subset was kept fixed, and validation data were derived only from the released training portion. Repeated training and paired prediction tests were not performed. The overlapping confidence intervals therefore support descriptive comparison rather than a definitive ranking of the five transformer families.

The study did not include same-protocol CNN or hybrid baselines, and no independent external cohort with compatible four-class labels, comparable input characteristics, and verifiable patient-level metadata was available for evaluation. Results from other datasets provide context but cannot replace direct testing under the same protocol.

Only two-dimensional images from the public benchmark were analyzed. The dataset provider describes all four classes as preprocessed T1-weighted contrast-enhanced MRI; however, the released JPEG/PNG files do not include DICOM acquisition metadata or image-level sequence tags, so sequence identity could not be independently verified for every image. Inter-slice continuity, volumetric tumor extent, and complementary MRI sequences were not represented. The broad random resized crop range may also alter lesion content or anatomical context, but this effect could not be quantified without lesion annotations.

Grad-CAM was assessed qualitatively because tumor masks, bounding boxes, expert localization ratings, and radiologist review were unavailable. Future evaluation should prioritize patient-level multicenter cohorts, repeated experiments, volumetric and multi-sequence inputs, anatomy-preserving augmentation, and analysis of annotation-supported explanations.

6. Conclusion

This study compared Vision Transformer, DeiT, Swin Transformer, BEiT, and EVA-02 for four-class brain tumor MRI classification under a common image-level protocol. BEiT-Base produced the highest observed accuracy of 0.9921 and macro F1-score of 0.9924 at a profiled operation count equal to that of the other patch-16 base configurations. The no-tumor class achieved a recall of 1.0000, but two glioma images assigned to this class reduced its precision to 0.9959. Overlapping Wilson 95% confidence intervals and mixed activation patterns do not support a decisive architecture ranking. These findings provide a within-transformer reference for the public two-dimensional benchmark, whose provider describes the released images as T1-weighted contrast-enhanced MRI. Confirmation in independent patient-level cohorts and repeated experiments is required before the results can be interpreted beyond this dataset.

Acknowledgments

The author acknowledges the providers of the publicly available Epic and CSCR Hospital Dataset through Mendeley Data.

Author Contributions

The author was responsible for the complete study, including conceptualization, methodology, software implementation, data organization, model training, validation, formal analysis, visualization, interpretation of results, manuscript writing, revision, and final approval of the submitted version.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The author declares that there are no competing financial or non-financial interests related to this work.

Data Availability Statement

The dataset is publicly available from Mendeley Data under the title Brain Tumor MRI Dataset (Glioma, Meningioma, Pituitary, No Tumor), also referred to as the Epic and CSCR Hospital Dataset (DOI: 10.17632/zwr4ntf94j.5). This study followed the published train-test split: the original test set was reserved for final evaluation, and the validation subset was derived only from the original training set. Patient and examination identifiers are not included; therefore, a patient-based split was not possible.

The model configurations, input resolution, augmentation settings, optimization hyperparameters, and evaluation metrics are documented in the manuscript. Source code, configuration files, and available trained weights can be obtained from the corresponding author upon reasonable request.

AI-Assisted Technologies Statement

During manuscript preparation, artificial intelligence-assisted language tools were used solely for translation, grammar and spelling correction, language refinement, and formatting. They were not used to generate data, design the methodology, conduct experiments, perform statistical analyses, select or modify results, or formulate scientific conclusions. All scientific content, analyses, interpretations, and final wording were critically reviewed and approved by the author, who assumes full responsibility for the manuscript.

References

1. Fornito A, Zalesky A, Breakspear M. The connectomics of brain disorders. Nat Rev Neurosci. 2015; 16: 159-172.
2. Weller M, Wen PY, Chang SM, Dirven L, Lim M, Monje M, et al. Glioma. Nat Rev Dis Primers. 2024; 10: 33.
3. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO classification of tumors of the central nervous system: A summary. Neuro Oncol. 2021; 23: 1231-1251.
4. Siegel RL, Kratzer TB, Wagle NS, Sung H, Jemal A. Cancer statistics, 2026. CA Cancer J Clin. 2026; 76: e70043.
5. Yadav AC, Kolekar MH. Deep feature-based approaches for brain tumor classification and segmentation in medical imaging. Biomed Signal Process Control. 2026; 117: 109603.
6. Ghorbian M, Ghorbian S, Ghobaei-Arani M. Advancements in machine learning for brain tumor classification and diagnosis: A comprehensive review of challenges and future directions. Arch Comput Methods Eng. 2026; 33: 1373-1408.
7. Pacal I, Ersoy M, Ozger F. DeformNeXt-Swin: A hybrid CNN-transformer framework for breast lesion classification in ultrasound and mammography. Chemometr Intell Lab Syst. 2026; 276: 105799.
8. Kunduracioglu I, Ince S, Bayram B, Kilicarslan S, Veziroglu E, Pacal I. Deep learning in acute ischemic stroke imaging: A systematic review of CT-and MRI-based segmentation, triage, and prognostic modeling. Neuroradiology. 2026. doi: 10.1007/s00234-026-04095-5.
9. Pacal I, Algarni A, Bayram B, Ince S. FA-UNet: A FasterNet and attention-gated hybrid network for precise ischemic stroke segmentation. J Integr Neurosci. 2025; 24: 40100.
10. Hayat M. Endoscopic image super-resolution algorithm using edge and disparity awareness. Bangkok, Thailand: Chulalongkorn University; 2023.
11. Hayat M, Izhar R, Nadeem M, Anjum H, Muhammad A, Bhattacharjee S. Hamsrnet-hybrid attention multiscale super-resolution network for endoscopic images. Proceedings of the 2025 5th International Conference on Digital Futures and Transformative Technologies (ICoDT2); 2025 December 17-18; Islamabad, Pakistan. New York, NY: IEEE.
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. Proceedings of the Ninth International Conference on Learning Representations; 2021 May 03-07. Available from: https://arxiv.org/abs/2010.11929.
13. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 October 10-17; Montreal, QC, Canada. New York, NY: IEEE.
14. Khan SUR, Asif S, Zhao M, Zou W, Li Y, Xiao C. ShallowMRI: A novel lightweight CNN with novel attention mechanism for Multi brain tumor classification in MRI images. Biomed Signal Process Control. 2026; 111: 108425.
15. Ul Ain N, Khan SA, Aladhadh S, Mir U, Ramzan M. Transformers meet CNNs: A comprehensive review and benchmarking of deep learning architectures for brain tumor classification in MRI. Comput Sci Rev. 2026; 60: 100897.

16. Hira MK, Hossain MS, Bithee MA, Sara U, Hasan MM, Towsif AA. Brain tumor MRI dataset (glioma meningioma pituitary no tumor): Version 5 [Internet]. Mendeley Data; 2025. Available from: https://data.mendeley.com/datasets/zwr4ntf94j/5.
17. Binish MC, RS SR, Thomas V. CBAM-SMK: Integrating convolution block attention module with separable multi-resolution kernels in deep neural networks for brain tumor classification. Biomed Signal Process Control. 2026; 112: 108483.
18. Hasan MJ, Hasan M, Akter S, Mahi AB, Uddin MP. Enhancing brain tumor classification with a novel attention based explainable deep learning framework. Biomed Signal Process Control. 2026; 112: 108636.
19. Meenal T, Asokan R. Quantum-inspired adaptive feature fusion for highly accurate brain tumor classification in MRI using deep learning. Biomed Signal Process Control. 2026; 112: 108694.
20. Sun J, Wen S, Tang C, Wu X, Wang S, Zhang Y. C-HGDAT: Hypergraph dynamic attention network with CNN-driven features for brain tumor classification. Biomed Signal Process Control. 2026; 112: 108747.
21. Tang Z, Liao X, Liao B, Shen C, Zhang Y. MRI brain tumor classification using RFENet three-branch model with SwishReLU. Biomed Signal Process Control. 2026; 113: 108893.
22. Wu H, Al-Huda Z. MDFE-Net: Multiscale dense feature extraction CNN for brain tumor classification from MRIs. Biomed Signal Process Control. 2026; 113: 108978.
23. Ke L, Hu G, Zhao M, Liu Z, Lv Z, Yang Y. Brain tumor classification from MRI images using a multi-scale channel attention CNN integrated with SVM. Sci Rep. 2026; 16: 6297.
24. Prayogo RD, Karimah SA, Nambo H. HayabusaNet: Hybrid attention-based multiscale fusion CNN for accurate and efficient brain tumor classification in MRI scans. Expert Syst Appl. 2026; 331: 133135.
25. Wang H, Zaqeem M, Fayaz M, Qiu D, Ahadzadeh S, Nguyen TN, et al. TumorNet: A hybrid lightweight framework for brain tumor classification and reasoning. Inf Sci. 2026; 746: 123423.
26. Ganie SM, Pacal I. DiSCNet: Directional split convolution for compute-efficient brain tumor diagnosis. Comput Biol Chem. 2026; 124: 109066.
27. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. Proceedings of the 38th International Conference on Machine Learning (ICML); 2021 July 18-24. pp. 10347-10357. Available from: https://proceedings.mlr.press/v139/touvron21a.
28. Bao H, Dong L, Wei F. BEiT: BERT pre-training of image transformers. Proceedings of the Tenth International Conference on Learning Representations (Virtual); 2022 April 25-29. Available from: https://arxiv.org/abs/2106.08254v1.
29. Fang Y, Sun Q, Wang X, Huang T, Wang X, Cao Y. Eva-02: A visual representation for neon genesis. Image Vis Comput. 2024; 149: 105171. doi: 10.48550/arXiv.2303.11331.
30. Pacal I. A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images. Int J Mach Learn Cybern. 2024; 15: 3579-3597.
31. Zhu H, Zhu Z, Lu SY. Multi-stage attention for efficient brain tumor classification with SAM-Med2D. Multimed Syst. 2026; 32: 51.
32. Ustun HI, Bulbul M, Yolcu Oztel G, Sahin VH. On-device brain tumor classification from MR images using smartphone. Adv Intell Syst. 2026; 8: 2500205.

33. Ducange P, Fazzolari M, Marcelloni F, Miglionico GC, Ruffini F. Fuzzy decision trees for explainable brain tumor classification: A comparative study with deep neural networks and classical binary decision trees. Inf Syst Front. 2026. doi: 10.1007/s10796-025-10683-2.
34. Fan Y, Wang X, Yue Z, Zhang X, Chen M, Chen J. Clustering-enhanced active learning with dynamic sampling for brain tumor classification. Biomed Signal Process Control. 2026; 118: 109715.
35. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 October 22-29; Venice, Italy. New York, NY: IEEE.