Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print) 2521-3520 (electronic)

ORIGINAL ARTICLE OPEN ACCESS

Brain Tumor Classification Using CNN and Grad-CAM on MRI Images

Rasha Jamal Hindi [a], and Fuat Türk [b]

[a] Computer Science, College of Education, Mustansiriyah University, Baghdad, Iraq
[b] Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara, Turkey

CORRESPONDENCE
Rasha Jamal Hindi
rashajamal94@uomustansiriyah.edu.iq

ARTICLE INFO
Received: Nov. 18, 2025
Revised: Feb. 20, 2026
Accepted: Feb. 28, 2026
Published: Mar. 30, 2026

ABSTRACT: ***Background:*** *Proper and interpretable brain tumor classification is essential in making a successful clinical decision in neuro-oncology. Automated approaches are potentially promising, but a lack of transparency in decision-making is usually an obstacle to clinical implementation.* ***Objective:*** *The proposed research developed and evaluated a convolutional neural network (CNN) model in the context of automatic brain tumor classification using magnetic resonance images (MRI) with a particular focus on a high-performing model and visualizing predictions using Gradient-weighted Class Activation Mapping (Grad-CAM).* ***Methods:*** *It utilized a dataset of 7,023 MRI scans as a sample, which was divided into glioma, meningioma, pituitary tumors, and no-tumor. Preprocessing of the data was done by normalizing and resizing, and stratifying into training, validation, and test subsets. The suggested CNN has been compared with the state-of-the-art transfer-learning architectures, such as VGG16, MobileNetV2, and DenseNet121.* ***Results:*** *The proposed CNN had the highest predictive accuracy of 94.75%, precision of 94.99%, recall of 94.75%, and an F1-score of 94.82%, and better than all the transfer-learning baselines. Moreover, Grad-CAM visualizations have always identified tumor-specific areas in the images, confirming the clinical plausibility of the model decisions.* ***Conclusions:*** *These results highlight the possibility of high-performance CNN-based classification used in conjunction with explainable AI to provide effective and high-quality diagnostic support that is accurate, dependable, and explainable by clinicians. The future research will explore the concept of multi-modal MRI integration, 3D architecture, and privacy-preserving deployment schemes in the context of real-life healthcare applications.*

KEYWORDS: *Convolutional Neural Networks; Magnetic Resonance Imaging; Brain Tumor Classification; Transfer Learning; Explainable Artificial Intelligence*

INTRODUCTION

Brain tumors (BTs) are among the most aggressive and life-threatening ailments globally, and this disease has a major implication on patient survival and quality of life. The development of tumors is a consequence of abnormal growth of cells, which form a mass or lesion that disrupts the normal functioning of the brain. The BT can be classified into benign and malignant types, depending on their nature, where the latter spreads very fast, leading to severe dysfunction without immediate intervention. The World Health Organization (WHO) gives some of the most common categories, which include meningiomas, gliomas, and pituitary tumors that are extremely variable in morphology in terms of location, texture, and size [1]. This type of heterogeneity, along with overlapping clinical manifestations including headaches and vision loss, aphasia, and cognitive deterioration, is especially problematic in terms of classification that is accurate and timely [2].

The use of MRI is the most effective in the detection and assessment of brain tumors because the technique provides the best resolution of soft-tissue contrast in a noninvasive fashion. MRI helps to detect structural abnormalities and to differentiate between tumors and normal tissues by radiologists [3]. Nevertheless, manual analysis of MRI scans is very labor-intensive and prone to error,

besides depending on the expertise of the radiologist. In addition, the growing amount of imaging data within contemporary medical settings, where medical images take up close to 90 percent of all clinical documentation, only increases the need to find automated, precise, and effective diagnostic interventions [4], [5]. Over the past few years, artificial intelligence (AI) and deep learning (DL) have become revolutionary technologies in medical imaging and have offered a significant improvement in diagnostic use in oncology, dermatology, and radiology [6], [7]. CNNs have proven to be very effective in feature extraction and classification. In the context of BT, they are often more effective than traditional machine learning methods [8], [9]. Despite these developments, there are two multifaceted issues that are still not addressed. To begin with, most of the high-performing models are computationally costly, including ensembles or hybrid transformer-based architectures, and hamper the scalability and clinical implementation [10]. Second, most deep learning models are black-box and hence lack interpretability, posing a question of trust, reliability and accountability by clinicians in making life-threatening decisions [7], [11]. As a way of overcoming these shortcomings, eXplainable Artificial Intelligence (XAI) has become popular in medical imaging studies [12], [13]. Other methods like Gradient-weighted Class Activation Mapping (Grad-CAM) offer visual heatmaps to mark tumor-relevant areas so clinicians can confirm and interpret model predictions in accordance with radiological appearance [14]. XAI techniques can improve the transparency of deep learning models by bridging the predictive performance and interpretability gap and promoting clinical trust and real-world adoption. Despite existing trends of using CNNs and Grad-CAM in medical imaging, the merit of this paper is that the explainability is placed and assessed in the context of the entire system design. This paper also considers interpretability as a design goal, as opposed to other studies that use Grad-CAM as an after-hoc visualization step; thus, its interpretation is not evaluated with respect to accuracy and efficiency. In particular, the proposed CNN is deliberately lightweight and parameter-efficient, without relying on computationally intensive techniques like ensembles, GAN-based augmentation pipelines, or transformer architectures, and also competes on a public benchmark dataset. Besides, it is not only with the help of Grad-CAM that correct predictions are visualized, but also facilitates qualitative inspection of doubtful cases or erroneous cases, which builds up clinical confidence and reliability. These challenges and opportunities inspire this paper to apply a simplified but efficient brain tumor classification system based on CNNs in combination with Grad-CAM. Compared to the previous methods that are focused on accuracy at the cost of transparency or computational speed, the proposed model focuses on having high predictive accuracy and interpretability. In particular, the following contributions are presented:

- A lightweight and parameter-efficient CNN model is developed to classify brain tumors into multiple categories using MRI images.
- Grad-CAM is integrated as a core component to identify discriminative tumor regions, providing visually interpretable explanations that support clinical understanding and verification.
- An extensive evaluation on a standard benchmark dataset is conducted, and the results are compared with standard transfer-learning baselines through the same experimental settings.
- The proposed framework represents the practical accuracy-interpretability-efficiency trade-off, which provides the ability of transparent decision support without the use of ensemble learning, GAN-based data augmentation, or transformer-based architectures.
- Grad-CAM is also utilized to study difficult and misclassified cases, which provides information about possible failure points and clarifies potential failure modes and model limitations.

The suggested model will combine high precision with explainability to provide a reliable AI-based brain tumor classification in clinical settings because it mitigates the twofold problem of diagnostic reliability and interpretability. The other parts of this paper are structured in the following way. Section II is a literature review of methods used to classify brain tumors, starting with earlier methods and proceeding to the more recent developments in deep learning and XAI. Section III describes the suggested framework in detail, which comprises the dataset, preprocessing pipeline, model architecture, training strategy, Explainability module, and evaluation metrics. Section IV presents the findings of the experiment and the comparison of the proposed approach to the state-of-the-art practices. Section V gives a detailed discussion of the findings, their implications, and limitations. Last but not least, Section VI presents the conclusion of the paper, which is the summation of the contributions and further research opportunities.

RELATED WORKS

Initial studies on brain tumor classification relied on handcrafted feature extraction combined with conventional machine learning classifiers. Raza *et al.* (2022) [15] marked an early transition toward deep learning by introducing DeepTumorNet, a hybrid CNN derived from GoogLeNet with modified layers and leaky ReLU activations, achieving strong benchmark performance and motivating wider adoption of CNN-based pipelines. As deep learning matured, several works continued to integrate traditional feature engineering and classical classifiers with explainability. Akter *et al.* (2023) [16] proposed a hybrid ML framework that extracted first-, second-, and higher-order statistical descriptors (including DWT) and applied dimensionality reduction with PCA, followed by multiple ML classifiers, reporting strong performance across accuracy, F1-score, and AUC-ROC while incorporating SHAP and LIME for interpretability. In the same year, Asiri *et al.* (2023) [17] introduced a dual-module approach combining image enhancement (adaptive Wiener filtering, neural networks, and ICA) with SVM-based segmentation and classification, reporting 98.9% accuracy and low processing time (0.43s), supporting feasibility for time-sensitive diagnostic settings. Also in 2023, Özkaraca *et al.* (2023) [18] proposed a modular Dense CNN that blends DenseNet, VGG16, and conventional CNN components; evaluated using both an 80/20 split and 10-fold cross-validation on a Kaggle setting, the model improved performance over standard transfer learning baselines at increased computational cost. Hossain *et al.* (2023) [19] further benchmarked multiple CNN families (VGG, ResNet, Inception, and Xception) and proposed an ensemble (IVX16) integrating the strongest candidates, reporting improved accuracy and demonstrating the practical benefits of combining complementary architectures. Simultaneously, lightweight models that foster interpretability started to appear: Iftikhar *et al.* (2023) [20] presented a simplified explainable CNN with reduced layers and parameters to decrease complexity without sacrificing performance on seen data and with interpretability through Grad-CAM, SHAP, and LIME. Narayankar and Baligar (2023) [21] specifically aimed to explain the use of a CNN on 7,043 Kaggle MRI images identified as a specific experiment; however, their results were less impressive, and the researchers considered several XAI methods (SHAP, LIME, Integrated Gradients, and Grad-CAM) to achieve an interpretability of the results on a region basis.

In 2024, subsequent works went in 3 primary directions, namely, better transfer learning, better explainability, and more architectural sophistication. A new paradigm called transfer-learning (Kumar et al, 2023) [22] created a better ResNet-50 both in benign vs malignant classification and showed high-quality results when compared to classical CNN baselines. Nazir *et al.* (2024) [23] introduced a tailored CNN on the BR35H dataset and used SHAP, LIME, and Grad-CAM to achieve interpretability, finding high validation rates and external dataset generalization. Hosny *et al.* (2024) [24] optimized ensemble transfer learning with DenseNet121 and InceptionV3 and modified fully connected layers, which demonstrated high accuracy and validated their selections with the help of Grad-CAM heatmaps. Roy *et al.* (2023) [25], [26] proposed an explainable ensemble model that combines DualGAN-based synthesis with ensemble feature extraction with Grad-CAM and achieves high accuracy with imbalanced data, respectively; Mandloi *et al.* (2023) [25], [26] followed this line and used a multi-phase method, conditional GAN augmentation, and LRP-based interpretability to achieve high accuracy with selected backbones. Also, Singh *et al.* (2024) [27] suggested a dynamic ensemble of CNN, ResNet-50, and EfficientNet-B5 with adaptive weighting and several XAI methods (Grad-CAM, SHAP, SmoothGrad, and LIME), which is highly accurate and can estimate uncertainty. Emerging hybrid and transformer-based approaches further increased performance and modeling capacity: Haque *et al.* (2024) [28] proposed NeuroNet19, a VGG19-based model enhanced with an inverted pyramid pooling module to capture multi-scale context, and interpreted predictions using LIME; Ahmed *et al.* (2024) [29] introduced a ViT–GRU hybrid to combine spatial and sequential representations and reported strong cross-validation performance with SHAP, LIME, and attention-based explanations; Islam *et al.* (2024) [30] proposed an EfficientNet-based framework achieving high accuracy on contrast-enhanced MRI. Alongside these, Nahiduzzaman *et al.* (2024) [31] developed a hybrid lightweight approach combining a parallel depthwise separable CNN with an RRELM classifier and CLAHE preprocessing, using SHAP for interpretability and reporting strong performance. More recent directions also include decentralized learning: Mastoi *et al.* (2024) [32] proposed a federated learning framework built on GoogLeNet to enable privacy-preserving training across multiple clients, incorporating Grad-CAM and saliency maps for interpretability.

The latest work in 2025 further reflects the continuing trend toward ensembles and enhanced explainability at scale. Sánchez-Moreno *et al.* (2025) [33] employed majority voting across VGG16, DenseNet121, and Inception-ResNet-v2 and provided interpretability using Grad–CAM++ and Integrated Gradients, reporting moderate multi-class accuracy in their evaluation setting. Overall, Table 1 provides a comparative overview of these approaches and illustrates the field's progression from ML-

based pipelines [16], [17] toward CNN architectures [15], [23], GAN-enhanced methods [25], [26], ensemble frameworks [27], and more recent transformer and federated-learning paradigms [29], [32]. Despite frequent reports of very high accuracy (often ≥98%), many state-of-the-art systems rely on computationally intensive designs such as ensembles, transformer-based models, or GAN-based augmentation, which can reduce reproducibility and increase deployment cost. In addition, it is frequently mentioned in the literature as a post-hoc analysis as opposed to an explicit design objective. The paper fills these gaps by proposing a simple yet efficient brain tumor classification system combining a lightweight CNN backbone with a Grad-CAM, with a focus on an accuracy-interpretability-efficiency trade-off. In this formulation, Grad-CAM serves not only as a visualization step but also as an integral component that supports clinically meaningful interpretation by highlighting tumor-relevant regions that align model decisions with radiological evidence.

Table 1. Comparative summary of recent brain tumor classification approaches

Ref.	Method or Framework	Dataset	Techniques	XAI Methods	Accuracy / Metrics	Key Contribution
[15]	DeepTumorNet (hybrid CNN)	Public MRI	Modified GoogLeNet	–	99.67% acc.; 100% recall	Custom hybrid CNN, Superior recall
[17]	Dual-module (enhancement + SVM)	Multi-type MRI	Wiener, ICA, NN + SVM	–	98.9% acc.; sens/spec. 0.991	Image enhancement + SVM, Fast inference (0.43 s)
[22]	Improved ResNet-50 (binary)	MRI (benign/ malignant)	Transfer learning	–	99.3% benign, 98.4% malignant	Binary diagnosis, TL outperforming CNN baselines
[18]	Modular Dense CNN	Kaggle MRI	DenseNet + VGG16 fusion	–	TL baselines	Modular design, Improved accuracy at higher cost
[25]	Ensemble (Dual-GAN + DeepEFF)	MRI (imbalanced)	GAN augmentation + ensemble	Grad-CAM	98.15% acc.	Class imbalance handling, Explainable ensemble
[26]	eGAN + DL + LRP	MRI (multi-class)	eGAN + CNN detection/classification	LRP	99.66% det.; 99.3% cls.	GAN-based balancing, Explainable pipeline
[19]	ViX16 (Ensemble ViT)	3,264 MRI	ViT ensemble	–	99.64%	Ensemble ViT benchmarking
[32]	CFLM (Federated GoogLeNet)	MRI (10 clients)	Federated learning	Grad-CAM, Saliency	94.4% acc.	Privacy-preserving learning
[21]	CNN + region-focused XAI	Kaggle MRI (7,043)	CNN baseline	LIME, SHAP, Grad-CAM	80% acc.	Interpretability-first analysis
[16]	Hybrid ML (statistical + ML classifiers)	Figshare/ Kaggle MRI	1st–3rd order stats, DWT, PCA	SHAP, LIME	Best across F1, AUC-ROC	Statistical texture features, ML-based classification, XAI-supported transparency
[23]	Customized CNN	BR35H (3,060)	Dataset-specific CNN	SHAP, LIME, Grad-CAM	100% train, 98.67% val.	High accuracy, Multi-XAI interpretability
[24]	Ensemble (Dense121 + IncV3)	Public MRI	TL with modified FC	Grad-CAM	99.02% acc.	Ensemble TL, Visual verification

Table 1. Continued

[28]	NeuroNet19 (VGG19 + IPPM)	Public (7k+)	Multi-scale pooling	–	99.3% acc.	Local/global feature capture
[29]	ViT–GRU hybrid	BrfTMHD-2023 + Kaggle	ViT + GRU	SHAP, LIME	98.97%	Spatial–temporal modeling, Interpretable hybrid
[30]	BrainNet (EfficientNet)	3,064 CE-MRI	EfficientNet scaling	–	99.69%	Efficient architecture, SOTA accuracy
[20]	Lightweight explainable CNN	MRI	Simplified CNN	Grad-CAM, SHAP, LIME	99.9% seen; 98.9% unseen	Lightweight design, Strong generalization
[31]	PSCNN–RRELM (hybrid)	MRI	CLAHE + depthwise CNN	SHAP, LIME	99.22% acc.	Lightweight hybrid, Local explanations
[27]	Dynamic ensemble	Benchmark MRI	CNN + ResNet + EfficientNet	Grad-CAM, SHAP, LIME	99.31% cross-dataset	Adaptive ensemble, Uncertainty-aware XAI
[33]	Ensemble (VGG16, Dense121, IRv2)	MRI (multi-class)	Majority voting	Grad-CAM++, IG	94.8% acc.	Robust ensemble, Detailed saliency analysis

MATERIALS AND METHODS

Figure 1 is the proposed structure that combines data preprocessing, a CNN design, and a module of explainability, using Grad-CAM, to obtain good and interpretable classification of brain tumors. The preprocessing of input MRI images is achieved by standardizing them by resizing, normalizing, and removing artifacts in order to provide consistency and enhance convergence of the models. The resulting processed images are fed into the CNN, which consists of sequential two-dimensional convolutional layers with the rectified linear unit (ReLU) activation to extract different levels of features hierarchically, interlaced with max-pooling layers that reduce the spatial dimensionality of the processed images, gradually leaving the salient tumor features. The resulting feature maps are flattened and then mapped with a dense layer, and later, it has a softmax output that categorizes each slice of the MRI to one of the four groups of glioma, meningioma, pituitary tumor, or no-tumor.

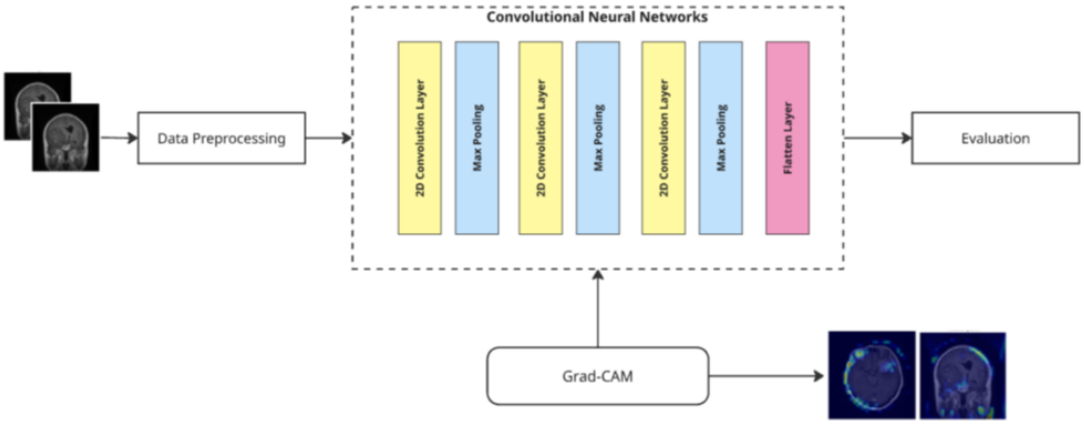

Figure 1. Proposed approach

Simultaneously, Grad-CAM is used on the last convolutional layer, which produces the class-discriminative saliency maps that identify the tumor-specific areas that contribute most to the classification choice. This two-fold design guarantees a strong predictive capability, as well as model transparency, since the emphasized heatmaps offer radiologically significant information that conforms to the tumor borders and increases clinical interpretability. Finally, the framework integrates both predictive accuracy and explainability, which is a major limitation of the traditional deep learning systems in medical imaging applications.

Dataset

This study was based on the Brain Tumor MRI Dataset, which consists of 7,023 MRI images with four classes: glioma, meningioma, pituitary tumor, and no-tumor. The dataset was obtained from a public repository (Kaggle) and is primarily composed of contrast-enhanced T1-weighted MRI scans. An independent test set of 1,311 images was reserved and used only for final evaluation; its class distribution is 300 glioma, 306 meningioma, 300 pituitary tumor, and 405 no-tumor.

To obtain a more robust estimate of model performance and reduce sensitivity to a single data partition, 5-fold stratified cross-validation was conducted on the remaining images (i.e., the non-test portion of the dataset). In each fold, the training portion was used to optimize model parameters, and a stratified validation subset was used for model selection and early stopping. The reported cross-validation results summarize the mean and standard deviation across the five folds, reflecting performance stability under different stratified splits. In addition to cross-validation, the held-out independent test set was kept fixed and was not used during training or validation in any fold, providing an unbiased final assessment of generalization.

Since the original images varied in quality and resolution, a standardized preprocessing pipeline was implemented, including conversion to RGB format, resizing to 128×128 pixels, and normalizing pixel values to the [0,1] range to ensure consistent input quality. On-the-fly data augmentation during training—random rotations, horizontal and vertical flipping, zooming, and brightness adjustments—was applied to reduce overfitting and improve generalization by addressing intraclass variability and imaging artifacts. Although mild class imbalance exists (e.g., 1,595 no-tumor and 1,321 glioma images), no aggressive rebalancing was performed because the distribution is not severely skewed; instead, stratification was applied in both the 5-fold cross-validation and the independent test split to preserve class proportions, and weighted precision, recall, and F1-scores were reported to ensure fair evaluation across classes. Finally, because the dataset originates from a single public source rather than multi-center or multi-protocol acquisitions, generalization to other clinical settings, MRI sequences, and institutions remains a limitation and should be validated using more diverse external datasets.

Figure 2 shows the class distribution of the Brain Tumor MRI dataset across the four categories (pituitary, no-tumor, meningioma, and glioma). The dataset is relatively balanced, with no-tumor images (1,595) slightly exceeding pituitary (1,457), meningioma (1,339), and glioma (1,321). This distribution reduces the risk of biased supervised learning; however, the remaining variations further support the use of stratified splitting to maintain fairness during training, validation, and testing.

Figure 3 depicts sample MRI scans within the no-tumor category with normal anatomy of the brain with no visible tumor masses. These samples bring out the common features of normal brain tissue, such as clearly defined ventricular areas and the absence of pathological proliferations. The heterogeneity of clinical MRI data is reflected by the variability of image quality and orientation, i.e., the difference between the coronal and the axial slices. This variety is useful to train deep learning models, as it provides better generalization capacity of the model to real-life situations, but also highlights the importance of such preprocessing operations as normalization, resizing, and augmentation to achieve resilience.

Preprocessing

Each MRI scan was processed using a standardized preprocessing pipeline to ensure consistency and reproducibility before model training. All images were resized to 128×128 pixels, converted to RGB format, and normalized to the [0,1] range to stabilize training and improve convergence.

To obtain a fair and balanced evaluation, the dataset was partitioned using a stratified splitting strategy that preserved class proportions (glioma, meningioma, pituitary tumor, and no-tumor) across all subsets. A held-out independent test set was reserved exclusively for final performance evaluation and was not used during training, validation, or cross-validation. The remaining data were used for model development, where 5-fold stratified cross-validation was conducted to assess robustness across

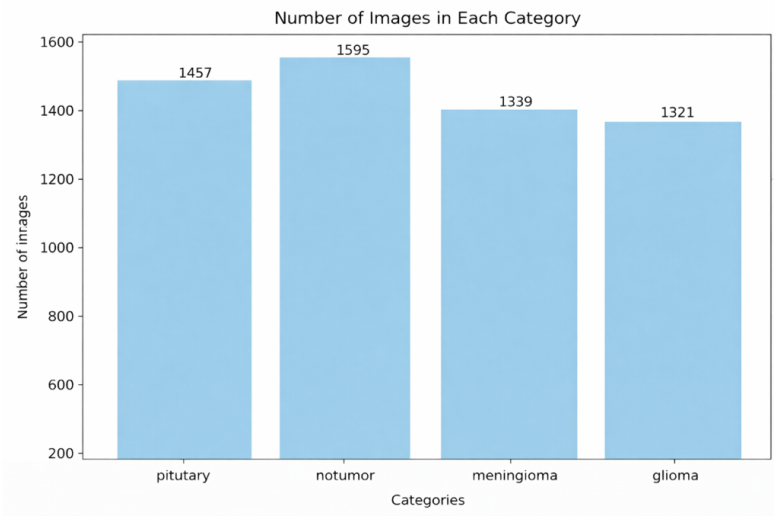

Figure 2. Distribution of images across categories

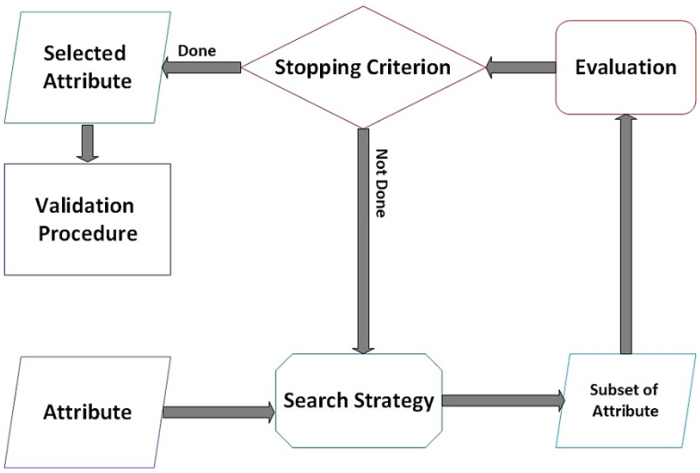

Figure 3. Sample MRI images from the No-Tumor class

different training and validation partitions. Within each fold, model parameters were optimized using the training subset, and hyperparameters were tuned with early stopping based on validation loss. A fixed random seed was used during data partitioning to ensure reproducibility of the experimental results.

Because the dataset does not provide patient-level identifiers, splitting at the patient level was not feasible; consequently, MRI slices from the same patient may appear in multiple subsets. This limitation may inflate the estimated generalization performance, as correlated samples in the training and test sets can reduce the apparent difficulty of the classification task. Future validation on datasets with explicit patient-level separation or multi-institutional cohorts will be necessary to confirm the robustness of the proposed model under stricter clinical evaluation protocols. To further improve the generalization and minimize overfitting, the training set was dynamically optimized with random rotations, horizontal and vertical flips, translations, and brightness modifications during training.

Model Architecture

A small CNN was used as the main classifier, and a number of transfer-learning baselines were used, such as VGG16, MobileNetV2, DenseNet121, and Xception, in order to allow a rigorous and fair comparative analysis. The proposed CNN processes standardized $128 \times 128 \times 3$ MRI slices with successive convolutional layers with 3×3 kernel and the rectified linear unit (ReLU) activation, which are interlaced with 2×2 max-pooling operations, which gradually reduce the spatial resolution and

enhance the representational depth respectively, using 64, 128, and 256 filters. The high-level feature maps are then flattened and fed to a fully connected layer of 512 units and dropout regularization to prevent overfitting, and then an output layer with 4 classes corresponding to glioma, meningioma, pituitary tumor, and no-tumor. Figure 4 provides a schematic view of the suggested CNN architecture, with the emphasis being made on the transition of convolutional feature extraction to dense classification and multi-class output.

To ensure a controlled and reproducible comparison between the proposed CNN and the transfer-learning baselines, all models were trained and evaluated using a unified experimental pipeline. Specifically, all experiments used the same preprocessing steps (RGB conversion, resizing to 128×128, and min–max normalization to the [0,1]range) and the same on-the-fly augmentation strategy applied only during training (random rotations, horizontal and vertical flipping, zooming, and brightness adjustments).

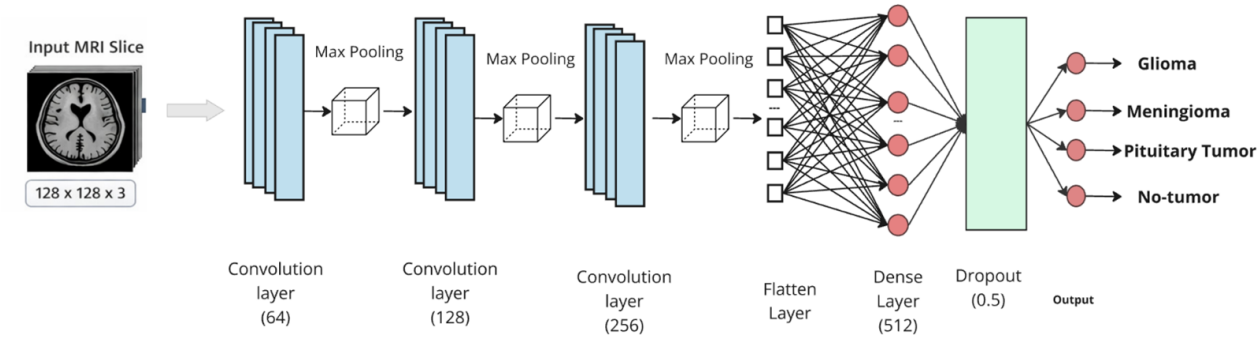

Figure 4. Architecture of the proposed CNN for multi-class brain tumor classification from MRI images

Evaluation was performed using two complementary protocols: (i) a fixed stratified split with a held-out independent test set reserved exclusively for final assessment, and (ii) 5-fold stratified cross-validation conducted on the remaining non-test portion of the dataset to quantify robustness across alternative training/validation partitions. In both protocols, stratification was used to preserve class proportions across subsets, and a fixed random seed was adopted to support reproducibility.

All models were optimized using categorical cross-entropy and the Adam optimizer with a batch size of 32 and an initial learning rate of 0.001, combined with a ReduceLROnPlateau scheduler. Training was run for up to 50 epochs, with early stopping applied consistently based on validation loss (patience = 5). Under the fixed split protocol, the best model checkpoint was selected using the validation set and then evaluated once on the independent test set. Under cross-validation, the same early-stopping criterion was applied within each fold, and results were aggregated across folds.

The only intentional methodological difference between models concerned initialization and adaptation strategy. VGG16 and DenseNet121 were initialized with ImageNet-pretrained weights and used primarily as feature extractors with a lightweight task-specific classification head, whereas MobileNetV2 and Xception additionally underwent partial fine-tuning of upper layers to improve domain adaptation. In contrast, the proposed CNN was trained from scratch (random initialization) while maintaining the same optimizer settings, training schedule, input resolution, and evaluation procedures.

In addition to these baselines, a CNN variant incorporating a spatial attention block was evaluated. In this configuration, an attention module was inserted after the final convolutional block to adaptively reweight spatial feature responses toward diagnostically relevant tumor regions before classification.

Finally, Grad-CAM was applied consistently across all architectures. Saliency maps were computed from the last convolutional layer to provide anatomically meaningful visual explanations by highlighting image regions that contributed most strongly to each predicted class, thereby supporting clinically interpretable model behavior and transparent decision-making.

Training Configuration and Hyperparameters

The configuration adopted in all the experiments, Table 2, is a summary of the training setup used in each experiment and was used in order to obtain consistency and a fair comparison between the different models evaluated. Any minor implementation-level variations are not expected to influence

the overall performance trends or the conclusions drawn in this study.

Table 2. Training configuration and hyperparameters used for all models

Component	Configuration
Input size	128 × 128 × 3
Convolution layers	3
Kernel size	3 × 3
Filters	64, 128, 256
Activation function	ReLU
Pooling	Max pooling (2 × 2)
Fully connected layer	512 units
Dropout rate	0.5
Loss function	Categorical cross-entropy
Optimizer	Adam
Initial learning rate	0.001
Learning-rate scheduler	ReduceLROnPlateau
Batch size	32
Maximum epochs	50
Early stopping	Enabled (patience = 5)
Weight initialization	ImageNet (for TL models)
Fine-tuning strategy	Partial fine-tuning of upper layers
Evaluation protocol	Best validation checkpoint / fixed-epoch

Gradient–weighted Class Activation Mapping (Grad-CAM)

To complement predictive performance with interpretability, an explainability module based on Grad-CAM was integrated into the proposed framework. Grad-CAM provides a class-discriminative localization map by highlighting the salient areas of the input MRI that most strongly influence the network's decision, thereby aligning the model's internal representation with clinically relevant tumor morphology.

$$\alpha_k^c = \frac{1}{u \times v} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{\partial y^c}{\partial A_{ij}^k}, \tag{1}$$

where $A^k \in \mathbb{R}^{u \times \nu}$denotes the k-th feature map of the final convolutional layer, and u and v represent its spatial dimensions (height and width), respectively. Here, A_{ij}^k is the activation at the spatial location (i, j), and y^c denotes the model output score for class c (before the SoftMax). Accordingly, α_k^c acts as an important weight that quantifies the contribution of the feature map k to class c. The Grad-CAM heatmap $L_{\text{Grad-CAM}}^c$ is then obtained by the weighted combination of feature maps, followed by a rectified linear unit (ReLU) to retain only positive influences:

$$L_{\text{Grad}-\text{CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{2}$$

The resulting heatmap is then upsampled to the resolution of the original MRI and overlaid as a color-coded map to provide a visual explanation of the classification outcome. Grad-CAM implementation is of particular significance to medical imaging since it contributes to filling the gap between the black-box deep learning models and the necessity of transparent and clinically valid decision support. Grad-CAM visualizations by validating the overlaps of the high-intensity activation areas with tumor-relevant ones make models more believable and aid radiological cross-validation. Additionally, it is a twofold benefit as this module enhances clinical trust in automated classification methods and

can be used as a diagnostic tool in the evaluation of possible misclassifications to make certain that predictions are accurate and interpretable.

Evaluation Metrics

In order to assess the model performance, a number of evaluation metrics, measured quantitatively was applied. The four measures were taken into account: accuracy, precision, recall, and Fl-score. Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The total accuracy, which is the percentage of the sample that was correctly classified, is given by (3).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

Nevertheless, the accuracy in itself might blur the class-wise differences, especially when the dataset is slightly imbalanced. To respond to this, the precision and recall were calculated on a per-class basis. Precision measures the proportion of the correctly predicted positive samples out of all the samples that are predicted to be positive, it is represented by (4).

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{4}$$

Where recall (or sensitivity) is the percentage of the genuine positives correctly recognized, it is defined by (5).

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{5}$$

The F1-score is the harmonic mean of both the precision and the recall and is a balanced measure that is resistant to class-based trade-offs. It is defined by (6).

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

To add strength, all measures were reported as weighted averages of classes, hence considering the proportion of each category in the final appraisal. This multi-metric assessment plan guarantees a quality and subtle characterization of the model performance beyond mere accuracy, which is essential in confirming clinical applicability towards the classification of brain tumors.

RESULTS AND DISCUSSION

The experimental results of the proposed CNN, which was trained and evaluated relative to transfer-learning baseline models, have been reported. The dataset was partitioned into training, validation, and independent testing as mentioned. In order to ensure a fair comparison, all of the models were trained using the same conditions, with a categorical cross-entropy loss function and Adam optimizer. The standard measures that have been used to perform the performance evaluation are accuracy, precision, recall, and the F1-score.

The main findings described in this section are based on two complementary evaluation protocols. First, a fixed stratified train/validation/test split was employed to remain consistent with widely used benchmark evaluation procedures on this publicly available dataset. A held-out independent test set was reserved exclusively for final performance assessment and was not used during training, validation, or cross-validation. Second, a 5-fold stratified cross-validation was conducted on the training/validation portion of the dataset to evaluate the robustness and stability of model performance under different data partitions.

A 5-fold cross-validation experiment was performed to assess performance variability across stratified folds. The proposed CNN achieved a mean classification accuracy of 94.75% with a standard deviation of ±0.16%, as shown in Table 3, indicating minimal variability across folds. Similar stability was observed for the transfer-learning baselines, suggesting consistent model behavior under repeated resampling. The low fold-wise variance demonstrates stable optimization dynamics and supports the generalizability of the learned feature representations.

In parallel, the independent stratified test split was used to provide an unbiased estimate of generalization performance on unseen data. Stratification ensured proportional class representation across training, validation, and test subsets, enabling fair and reliable performance evaluation. In this work,

Table 3. Results of 5-fold cross-validation for the proposed CNN model, with the mean and standard deviation of accuracies for each fold

Fold	Proposed CNN	DenseNet121	VGG16	MobileNetV2
1	94.54%	93.16%	90.41%	90.37%
2	94.81%	93.29%	90.34%	90.14%
3	94.80%	92.93%	90.35%	90.27%
4	94.94%	92.96%	90.05%	90.02%
5	94.64%	93.16%	90.37%	90.33%
Mean±Std	94.75±0.16%	93.10±0.15%	90.30±0.15%	90.23±0.14%

a fixed stratified split was used to separate the data into training, validation, and a held-out independent test set. This approach ensures a consistent benchmark-style evaluation and provides an unbiased estimate of generalization performance on unseen data. Stratification preserves the class distribution across all subsets, which is essential for fair and reliable model assessment.

In addition to the fixed split evaluation, 5-fold stratified cross-validation was conducted on the training/validation portion of the dataset to assess the robustness and stability of model performance under different data partitions. Cross-validation complements the independent test evaluation by quantifying variability across folds and reducing the likelihood that results depend on a single favorable split.

Accordingly, the reported performance metrics include both independent test set results and cross-validation statistics, providing a comprehensive and reliable evaluation of the proposed framework.

The CNN model training curves depicted in Figure 5 have both correct and steady convergence rates in both the accuracy and loss metrics. The training accuracy exhibits a sharp increase in the initial epochs, with a value of over 95 percent in the 7th epoch, and plateaus at 98 percent in the 10th epoch. Validation accuracy reached 94%, indicating good generalization with limited overfitting. In line with this, the training loss keeps decreasing over the epochs, and the corresponding validation loss also continues the same pattern, approaching 0.25 with slight fluctuations. The fact that training and validation performance are almost identical indicates the strength of the model and indicates that the regularization methods implemented, including dropout, were effective in counteracting overfitting. These findings indicate that the CNN has been able to learn discriminative features using the MRI scans with high predictive accuracy, with a stable optimization dynamic.

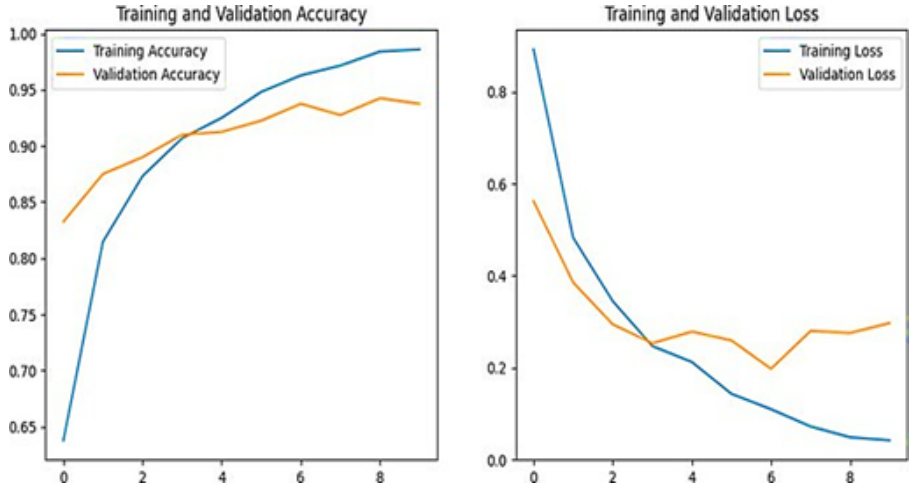

Figure 5. Accuracy and Loss Curves of the Proposed Custom CNN

Table 4 reports performance on the held-out independent test set. Model selection was performed using the best validation checkpoint during training, and the test set was not involved in any stage of training or cross-validation. The proposed custom CNN achieves the strongest performance among the evaluated models, attaining an accuracy of 94.75% and an F1-score of 0.9482 on the independent test set. This result confirms the effectiveness of a task-specific lightweight architecture in extracting

discriminative features from brain tumor MRI scans while maintaining computational simplicity and efficiency.

Table 4. Performance comparison of CNN and transfer-learning models on the Brain Tumor MRI Dataset (best validation-checkpoint)

Model	Accuracy	Precision	Recall	F1 Score
CNN	0.9475	0.9499	0.9475	0.9482
DenseNet121	0.9325	0.9328	0.9325	0.9321
VGG16	0.9050	0.9116	0.9050	0.9066
MobileNetV2	0.9025	0.9098	0.9025	0.9013

DenseNet121 follows with an accuracy of 93.25%, likely benefiting from dense feature reuse and strong representational capacity. VGG16 and MobileNetV2 achieve stable but comparatively lower accuracies under the same preprocessing, augmentation, and training conditions. Across all evaluated architectures, consistently high precision values indicate effective reduction of false positives, while recall values demonstrate balanced sensitivity across tumor classes. Overall, these results highlight the competitive generalization ability of the proposed CNN on unseen data while preserving architectural simplicity and interpretability.

Several studies in Table 5 report markedly higher accuracies (often >98%); however, these values are typically obtained under experimental conditions that differ substantially from the present work, including the use of alternative datasets, multi-source acquisitions, aggressive GAN-based augmentation, ensemble or transformer-based architectures, multi-modal MRI inputs, and/or evaluation strategies such as k-fold cross-validation. Such methodological differences limit the fairness of direct numerical comparison and can also introduce additional sources of variability that are not always transparent or easily reproducible.

Table 5. Comparison of the proposed CNN with recent methods reporting lower accuracy, highlighting the use of XAI

Ref.	Method	Accuracy (%)	Dataset / Evaluation Setting	Notes on Comparability
[21]	CNN baseline + XAI	80.00	Kaggle MRI (~7,043 images)	Same benchmark family; shallow baseline
[33]	Ensemble (VGG16, Dense121, IRv2) + XAI	86.17	MRI multi-class (benchmark-style)	Ensemble-based; different training setup
[17]	Enhancement + SVM pipeline	98.90	Different public MRI dataset	Different dataset; non-CNN classifier
[23]	Customized CNN + multi-XAI	98.67 (val)	BR35H dataset (3,060 images)	Different dataset and validation protocol
[15]	Hybrid CNN (DeepTumorNet)	99.67	Public MRI datasets	Hybrid architecture; different data sources
[25]	GAN + ensemble + Grad-CAM	98.15	Imbalanced MRI datasets	GAN-based augmentation and ensemble learning
[29]	ViT–GRU hybrid + XAI	98.97	Multi-source MRI; 10-fold CV	Transformer-based; cross-validation
[30]	EfficientNet-based framework	99.69	CE-MRI dataset	Different dataset; optimized scaling
[24]	Ensemble TL + Grad-CAM	99.02	Public MRI datasets	Ensemble transfer learning
Proposed method	Custom CNN + Grad-CAM	94.75	Kaggle MRI (7,023 images)	Lightweight, interpretable

Importantly, the objective of this study is not solely to maximize accuracy, but to demonstrate a practical accuracy–interpretability–efficiency trade-off suitable for clinical decision support. In neuro-oncological imaging, model adoption depends not only on predictive performance but also on ex-

plainability, reliability, and deployability. The proposed approach achieves competitive accuracy on a single public benchmark using a lightweight, parameter-efficient CNN while integrating Grad-CAM to provide visually interpretable evidence that predictions are driven by tumor-relevant regions. This interpretability is a clinically significant feature as it allows qualitative testing of model behavior and facilitates the analysis of errors and minimizes the possibility of a reliance on spurious image features that remain undetected.

In addition, methods with very high accuracy can also be less costly in terms of significantly greater computational complexity, larger memory footprints, longer training and inference times, and lower transparency (e.g., ensembles and transformer-based models). Such aspects can cause impediments to reproducibility and restrict its use in resource-constrained settings. In comparison, the suggested model focuses on computational efficiency and explainability without significantly affecting classification performance, so it is more suitable for the conditions of the real world, where inference timeliness, model transparency, and interpretability are of significant concern to clinician trust. Therefore, when there is a minor decline in peak accuracy with accompanying enhancement in interpretability and reduction in the computational load, it can be tolerated, especially when the model is to be provided as a decision-support tool rather than a standalone diagnostic system.

Together with the aggregate performance metrics, the qualitative per-class error analysis was performed according to the analysis of the prediction results and visualizations based on Grad-CAM. The findings suggest that the no-tumor class is the least uncertain category of those who are identified with a distinct separation between the tumor classes because of the lack of abnormal contrast-enhanced areas. Among tumor categories, glioma and meningioma present the greatest classification challenge, which can be attributed to overlapping intensity patterns and spatial characteristics in contrast-enhanced T1-weighted MRI scans. Gliomas, in particular, exhibit heterogeneous morphology and infiltrative growth patterns, increasing the likelihood of confusion with adjacent tumor classes.

Although a detailed confusion matrix and explicit per-class precision–recall statistics are not presented here, the observed qualitative error patterns align with previously reported findings in brain tumor classification studies using comparable public datasets. These trends highlight the intrinsic difficulty of distinguishing tumor subtypes with partially overlapping radiological features. Future work will incorporate a comprehensive quantitative class-wise analysis, including confusion matrices and sensitivity-specificity breakdowns, to further characterize misclassification behavior and strengthen robustness assessment.

Figure 6 of the Grad-CAM visualizations shows that the model is able to localize tumor regions in both cases of glioma and meningioma in a clinically significant way. In the glioma case (top row), the heatmap highlights the abnormal mass in the middle of the brain, well-fitting the pathological lesion in the MRI, and also highlights the surrounding tissue that is indicative of infiltrative growth that is characteristic of gliomas. The activation map in the meningioma case (bottom row) clearly indicates the abnormal growth on the right hemisphere, whereby the saliency is high in the boundaries of the tumors, which are well aligned with the radiological characteristics of meningioma. The locality of the foci of interest without the use of overactivation of irrelevant brain areas suggests that the network relies on tumor-related features to make its predictions, as opposed to spurious correlations. These findings support the claim that Grad-CAM integration not only enhances the interpretability of the CNN but also offers radiologically sensible reasons as to why the CNN arrived at its decision, which supports its appropriateness as a clinically reliable decision support system.

Extended Epoch Analysis

To characterize the effect of training duration on convergence and generalization, all models were additionally evaluated under a fixed-epoch protocol at 20 and 50 epochs using the same preprocessing, augmentation, and optimization settings. This analysis addresses a different methodological objective than the checkpoint-based results reported in Table 4. Specifically, Table 4 reports performance from the best validation checkpoint (i.e., model selection based on validation loss under early stopping), whereas Table 6 reports performance after completing a predefined number of epochs without model selection. Because these protocols differ in how the final model state is determined, the resulting metrics are not expected to coincide and should not be interpreted as directly comparable estimates of peak model performance.

As shown in Table 6, model behavior under extended training varies by architecture. Xception exhibits strong performance at both 20 and 50 epochs, consistent with its high-capacity pretrained backbone and depthwise separable convolutional design, which can benefit from prolonged optimization. MobileNetV2 also benefits as the number of training epochs increases, suggesting that lightweight pretrained models can take more training epochs to fully fit the target domain. Conversely, the

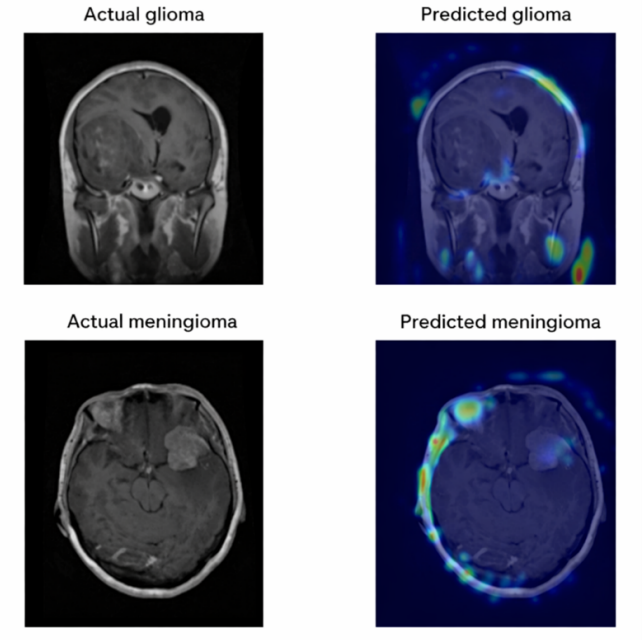

Figure 6. Grad-CAM visualizations for glioma and meningioma cases

suggested compact CNN demonstrates a minor reduction between 20 and 50 epochs. The CNN with attention enhances slightly during a long training period, indicating that attention-based re-weighting could also be used to improve the use of features with a long training duration. Lastly, InceptionV3 is slightly decreasing to 50 epochs, and this could possibly indicate an additional risk of overfitting in higher capacity models when trained on a single-source dataset with a low degree of diversity. Importantly, the fixed-epoch results indicate that some pretrained architectures (notably Xception and MobileNetV2) can achieve higher accuracy under prolonged training in this evaluation setting. However, fixed-epoch performance reflects a combined effect of architectural capacity and training duration rather than a controlled model-selection comparison. In this work, the primary objective is not to optimize for absolute peak accuracy under extended training, but to establish a deployable accuracy–interpretability–efficiency trade-off. The proposed compact CNN achieves competitive performance with reduced architectural complexity and provides consistent Grad-CAM explanations that support clinical verification of model behavior. In practical deployment, such transparency and computational efficiency can be critical alongside high accuracy, particularly in resource-constrained environments where interpretability and predictable failure modes are required for clinician trust.

Table 6. Performance comparison of models trained on the brain tumor MRI dataset (fixed-epoch evaluation at 20 and 50 epochs)

Model	**20 Epochs**				**50 Epochs**			
	Accuracy	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Xception	0.9786	0.9786	0.9786	0.9786	0.9822	0.9824	0.9822	0.9822
CNN	0.9644	0.9641	0.9644	0.9642	0.9626	0.9631	0.9626	0.9627
MobileNetV2	0.9644	0.9644	0.9644	0.9644	0.9786	0.9792	0.9786	0.9788
CNN & Attention	0.9609	0.9609	0.9609	0.9608	0.9680	0.9680	0.9680	0.9676
InceptionV3	0.9448	0.9480	0.9448	0.9456	0.9359	0.9389	0.9359	0.9367

Statistical Significance Analysis

To evaluate whether the observed performance differences between the proposed CNN and the DenseNet121 baseline are statistically meaningful rather than attributable to random variation, both prediction-level and fold-level statistical significance analyses were conducted. These complementary tests assess differences in paired predictions on the same test samples as well as consistency across

cross-validation folds.

The paired classification results on the held-out independent test set were first subjected to McNemar test, where both models were tested on the same samples. This non-parametric test is used to determine whether the two classifiers have significantly different error patterns when they are used under identical conditions. The 2×2 contingency table is presented in Table 7.

Table 7. McNemar's Test

	DenseNet121 correct	**DenseNet121 incorrect**
CNN correct	1149	97
CNN incorrect	62	3

Second, to assess the persistence of the performance gap with different data partitions, a paired t -test was conducted on the five cross-validation fold accuracies of the two models. Since CNN and DenseNet121 are tested on the same folds, a paired design is the right choice to correct the variability of the folds. The test provided t=14.02 and p=0.0002, which means that the difference between the mean accuracy of folds is statistically significant.

Besides testing the hypothesis, the size of the difference was also measured with the help of Cohen's d, which was 10.69, which is a very large effect size according to traditional standards. Together, the prediction-level McNemar test (independent test set) and the fold-level paired t-test with effect-size estimation (cross-validation) give converging indications that the proposed CNN has consistently higher classification performance compared to DenseNet121 under identical experimental conditions.

CONCLUSION

This paper presented a CNN to classify BT based on MRI images, and the Grad-CAM module was incorporated to make the results more understandable. The proposed CNN demonstrated better results in terms of performance than the popular models of transfer-learning based on VGG16, MobileNetV2, and DenseNet121, in the course of massive testing on a benchmark dataset. The model achieved high accuracy, precision, recall, and F1-scores, proving that it can be robust in discriminating between glioma, meningioma, pituitary tumors, and non-tumor cases. Notably, the Grad-CAM images demonstrated that the model made decisions based on clinically meaningful tumor locations, hence closing the gap between predictive and explainable methods. The findings point to two major contributions: first, it was shown that a well-shaped CNN can outperform deeper and more parameter-intensive models in domain-specific medical imaging tasks; second, it has been shown that Grad-CAM can be used as a validation method that serves as a transparent and reliable way of understanding what a model predicts. These results support the importance of deep learning as an effective diagnostic tool but also as a decision-support system with a high level of clinical reliability. The future areas of work will include the development of the framework to consider multi-modal MRI input schemes, consider three-dimensional CNN and transformer-based design, and test the system on larger and multi-institutional datasets to further prove its applicability. Also, problems of federated learning and privacy-protective methods will be explored to help to implement safe implementation in the real clinical setting. Altogether, the study is a contribution to the increasing amount of literature that deep learning, combined with Explainability, can make a significant step towards improving computer-aided diagnosis in neuro-oncology.

SUPPLEMENTARY MATERIAL

None.

AUTHOR CONTRIBUTIONS

Rasha Jamal Hindi: Conceptualization, Methodology, Investigation, and Writing - review & editing. Fuat Türk: Formal analysis, and Investigation.

FUNDING

This research received no external funding.

DATA AVAILABILITY STATEMENT

The dataset used and analyzed during the current study, "Brain Tumor MRI Dataset", is publicly available on Kaggle at: (https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset). Additional processed data are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Gazi University and Mustansiriyah University for their support and for providing the necessary facilities that made this research possible.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Grammarly for grammar checking and language polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, *et al.*, "The 2021 WHO classification of tumors of the central nervous system: A summary," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021, doi: 10.1093/neuonc/noab106.

[2] Q. T. Ostrom, G. Cioffi, K. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018," *Neuro-Oncology*, vol. 23, no. Supplement_3, pp. iii1–iii105, 2021, doi: 10.1093/neuonc/noab200.

[3] Y. Chen, C.-B. Schönlieb, P. Liò, T. Leiner, P. L. Dragotti, G. Wang, D. Rueckert, D. Firmin, and G. Yang, "AI-based reconstruction for fast MRI—A systematic review and meta-analysis," *Proceedings of the IEEE*, vol. 110, no. 2, pp. 224–245, 2022, doi: 10.1109/jproc.2022.3141367.

[4] J. Huang, Y. Fang, Y. Wu, H. Wu, Z. Gao, Y. Li, J. D. Ser, J. Xia, and G. Yang, "Swin transformer for fast MRI," *Neurocomputing*, vol. 493, pp. 281–304, Jul. 2022, doi: 10.1016/j.neucom.2022.04.051.

[5] Y. Nan, J. D. Ser, S. Walsh, C. Schönlieb, M. Roberts, I. Selby, K. Howard, J. Owen, J. Neville, J. Guiot, *et al.*, "Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions," *Information Fusion*, vol. 82, pp. 99–122, Jun. 2022, doi: 10.1016/j.inffus.2022.01.001.

[6] B. H. Kann, A. Hosny, and H. J. Aerts, "Artificial intelligence for clinical oncology," *Cancer Cell*, vol. 39, no. 7, pp. 916–927, 2021, doi: 10.1016/j.ccell.2021.04.002.

[7] S. G. Anjara, A. Janik, A. Dunford-Stenger, K. Mc Kenzie, A. Collazo-Lorduy, M. Torrente, L. Costabello, and M. Provencio, "Examining explainable clinical decision support systems with think aloud protocols," *PLOS ONE*, vol. 18, no. 9, Art no. e0291443, 2023, doi: 10.1371/journal.pone.0291443.

[8] M. A. Gómez-Guzmán, L. Jiménez-Beristaín, E. E. García-Guerrero, O. R. López-Bonilla, U. J. Tamayo-Perez, J. J. Esqueda-Elizondo, K. Palomino-Vizcaino, and E. Inzunza-González, "Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks," *Electronics*, vol. 12, no. 4, Art no. 955, 2023, doi: 10.3390/electronics12040955.

[9] R. A. Zeineldin, M. E. Karar, Z. Elshaer, J. Coburger, C. R. Wirtz, O. Burgert, and F. Mathis-Ullrich, "Explainability of deep neural networks for MRI analysis of brain tumors," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 9, pp. 1673–1683, 2022, doi: 10.1007/s11548-022-02619-x.

[10] R. A. Zeineldin, M. E. Karar, Z. Elshaer, J. Coburger, C. R. Wirtz, O. Burgert, and F. Mathis-Ullrich, "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI," *Scientific Reports*, vol. 14, no. 1, Art no. 3713, 2024, doi: 10.1038/s41598-024-54186-7.

[11] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, pp. 1–85, 2022, doi: 10.1214/21-ss133.

[12] E. Tjoa and C. Guan, "Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 858–870, 2023, doi: 10.1109/tai.2022.3228834.

[13] D. Bhati, F. Neha, and M. Amiruzzaman, "A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging," *Journal of Imaging*, vol. 10, no. 10, Art no. 239, 2024, doi: 10.3390/jimaging10100239.

[14] E. Tjoa, H. J. Khok, T. Chouhan, and C. Guan, "Enhancing the confidence of deep learning classifiers via interpretable saliency maps," *Neurocomputing*, vol. 562, Art no. 126825, Dec. 2023, doi: 10.1016/j.neucom.2023.126825.

[15] A. Raza, H. Ayub, J. A. Khan, I. Ahmad, A. S. Salama, Y. I. Daradkeh, D. Javeed, A. Ur Rehman, and H. Hamam, "A hybrid deep learning-based approach for brain tumor classification," *Electronics*, vol. 11, no. 7, Art no. 1146, 2022, doi: 10.3390/electronics11071146.

[16] S. Akter, M. Simul Hasan Talukder, S. K. Mondal, M. Aljaidi, R. Bin Sulaiman, and A. A. Alshammari, "Brain tumor classification utilizing pixel distribution and spatial dependencies higher-order statistical measurements through explainable ML models," *Scientific Reports*, vol. 14, no. 1, Art no. 25800, 2024, doi: 10.1038/s41598-024-74731-8.

[17] A. A. Asiri, T. A. Soomro, A. A. Shah, G. Pogrebna, M. Irfan, and S. Alqahtani, "Optimized brain tumor detection: A dual-module approach for MRI image enhancement and tumor classification," *IEEE Access*, vol. 12, pp. 42 868–42 887, 2024, doi: 10.1109/access.2024.3379136.

[18] O. Özkaraca, O. İ. Bağrıaçık, H. Gürüler, F. Khan, J. Hussain, J. Khan, and U. E. Laila, "Multiple brain tumor classification with dense CNN architecture using brain MRI images," *Life*, vol. 13, no. 2, Art no. 349, 2023, doi: 10.3390/life13020349.

[19] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, "Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1261–1272, 2024, doi: 10.1109/jbhi.2023.3266614.

[20] S. Iftikhar, N. Anjum, A. B. Siddiqui, M. Ur Rehman, and N. Ramzan, "Explainable CNN for brain tumor detection and classification through XAI based key features identification," *Brain Informatics*, vol. 12, no. 1, Art no. 10, 2025, doi: 10.1186/s40708-025-00257-y.

[21] P. Narayankar and V. P. Baligar, "Explainability of brain tumor classification based on region," in *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, IEEE, Apr. 2024, 1–6, doi: 10.1109/icetcs61022.2024.10544289.

[22] S. Kumar, S. Choudhary, A. Jain, K. Singh, A. Ahmadian, and M. Y. Bajuri, "Brain tumor classification using deep neural network and transfer learning," *Brain Topography*, vol. 36, no. 3, pp. 305–318, 2023, doi: 10.1007/s10548-023-00953-0.

[23] M. I. Nazir, A. Akter, M. A. Hussen Wadud, and M. A. Uddin, "Utilizing customized CNN for brain tumor prediction with explainable AI," *Heliyon*, vol. 10, no. 20, Art no. e38997, 2024, doi: 10.1016/j.heliyon.2024.e38997.

[24] K. M. Hosny, M. A. Mohammed, R. A. Salama, and A. M. Elshewey, "Explainable ensemble deep learning-based model for brain tumor detection and classification," *Neural Computing and Applications*, vol. 37, no. 3, pp. 1289–1306, 2024, doi: 10.1007/s00521-024-10401-0.

[25] P. Roy, F. M. S. Srijon, and P. Bhowmik, "An explainable ensemble approach for advanced brain tumor classification applying Dual-GAN mechanism and feature extraction techniques over highly imbalanced data," *PLOS ONE*, vol. 19, no. 9, Art no. e0310748, 2024, doi: 10.1371/journal.pone.0310748.

[26] S. Mandloi, M. Zuber, and R. K. Gupta, "An explainable brain tumor detection and classification model using deep learning and layer-wise relevance propagation," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 33 753–33 783, 2023, doi: 10.1007/s11042-023-16708-9.

[27] R. Singh, S. Gupta, A. O. Ibrahim, L. A. Gabralla, S. Bharany, A. U. Rehman, and S. Hussen, "Advanced dynamic ensemble framework with explainability driven insights for precision brain tumor classification across datasets," *Scientific Reports*, vol. 15, no. 1, Art no. 29090, 2025, doi: 10.1038/s41598-025-14917-w.

[28] R. Haque, M. M. Hassan, A. K. Bairagi, and S. M. Shariful Islam, "NeuroNet19: An explainable deep neural network model for the classification of brain tumors using magnetic resonance imaging data," *Scientific Reports*, vol. 14, no. 1, Art no. 1524, 2024, doi: 10.1038/s41598-024-51867-1.

[29] M. M. Ahmed, M. M. Hossain, M. R. Islam, M. S. Ali, A. A. N. Nafi, M. F. Ahmed, K. M. Ahmed, M. S. Miah, M. M. Rahman, M. Niu, *et al.*, "Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh," *Scientific Reports*, vol. 14, no. 1, Art no. 22797, 2024, doi: 10.1038/s41598-024-71893-3.

[30] M. M. Islam, M. A. Talukder, M. A. Uddin, A. Akhter, and M. Khalid, "BrainNet: Precision brain tumor classification with optimized EfficientNet architecture," *International Journal of Intelligent Systems*, vol. 2024, no. 1, Art no. 3583612, 2024, doi: 10.1155/2024/3583612.

[31] M. Nahiduzzaman, L. F. Abdulrazak, H. B. Kibria, A. Khandakar, M. A. Ayari, M. F. Ahamed, M. Ahsan, J. Haider, M. A. Moni, and M. Kowalski, "A hybrid explainable model based on advanced machine learning and deep learning models for classifying brain tumors using MRI images," *Scientific Reports*, vol. 15, no. 1, Art no. 1649, 2025, doi: 10.1038/s41598-025-85874-7.

[32] Q. Mastoi, S. Latif, S. Brohi, J. Ahmad, A. Alqhatani, M. S. Alshehri, A. Al Mazroa, and R. Ullah, "Explainable AI in medical imaging: An interpretable and collaborative federated learning model for brain tumor classification," *Frontiers in Oncology*, vol. 15, Feb. 2025, doi: 10.3389/fonc.2025.1535478.

[33] L. Sánchez-Moreno, A. Perez-Peña, L. Duran-Lopez, and J. P. Dominguez-Morales, "Ensemble-based convolutional neural networks for brain tumor classification in MRI: Enhancing accuracy and interpretability using explainable AI," *Computers in Biology and Medicine*, vol. 195, Art no. 110555, Sep. 2025, doi: 10.1016/j.compbiomed.2025.110555.