



Enhancing Brain Tumor Classification with Explainable AI: A Comparative Study of CNN Model

Vishvajet Choudhary¹, Yogita D. Patil^{1,*}, Karthik Kurup¹, Tanishka Gophane¹
, Mahesh Kapse¹ and Pushkar Kumar¹

¹ Vishwakarma University, Pune, India
yogita.patil1@vupune.ac.in*

Abstract. Classification of brain tumor from MRI is among the foremost important processes within the healthcare industry because it facilitates early diagnosis of the patients, which further improves their quality of life. The objective is to investigate the application of deep learning methods using the CNN architecture along with the XAI for brain tumors classification in multi-class. In this regard, transfer learning technique has been adopted in this research work in order to employ four pre-trained deep learning models known as InceptionV3, EfficientNetB0, DenseNet121, and ResNet50. These models have been tested by using a balanced dataset of MRI images which included glioma, meningioma, pituitary tumor, and no tumor with a total number of 7,200 images. Feature extraction and fine-tuning were used for the training of all four models within the target domain. It was observed that ResNet50 yielded the most desirable outcomes among the four tested models with the highest training accuracy rate of 98.09% and test accuracy of 92.06% alongside precision, recall, and F1-score. To solve the problem of interpretability of deep learning models, Grad-CAM technique has been applied to the model having the best performances. This work represents an effort towards developing efficient and interpretable AI-powered diagnostic systems for the diagnosis of brain tumors.

Keywords: Medical Image Analysis, MRI, Brain Tumor, Deep Learning, CNN, ResNet50, EfficientNet, DenseNet, Inception V3, XAI.

1 Introduction

The existence of brain tumors is among the major challenges faced within the field of neuro-oncology mainly due to the very high mortality rate related to these diseases and their complicated anatomical structure as well. The development of brain tumors is as a result of uncontrolled cell growth within the brain, which hinders brain performance and which poses serious risks to the patient's life in case of their non-detection in time [1]. According to WHO, brain tumors may be classified into benign and malignant brain tumors based on the disease's morphologic and molecular features, where glioblastoma stands as a kind of malignant brain tumors [3].

It is now considered that MRI is a recommended imaging technique for diagnosing brain tumors due to the high-resolution capabilities of imaging the soft tissues, together with the production of high-resolution anatomical images without exposing patients to any kind of radiation [7]. In addition, various MRI sequences (T1-, T2-weighted, and FLAIR images) offer extra information about the tumor margins, as well as edema and structural abnormalities [7]. However, MRI interpretation performed by radiologists is sometimes labor-intensive, subjective and prone to variation, especially when classifying multiclass brain tumors [10].

Currently, AI and particularly DL are gaining ground in the domain of medical image analysis. The hierarchical feature extraction automatically in MRI images through the CNNs has lessened the need for manually extracting features, making the process much faster [6]. DL architectures like ResNet, DenseNet, EfficientNet, and Inception have proven themselves in their ability to achieve high classification accuracies in distinguishing between brain tumor images, obtaining accuracies of above 95% [2], [4], [37]. They use a method known as transfer learning, which is used to overcome the challenge of insufficient labeled data [11].

One of the major difficulties in deep learning approaches is interpretation. Traditionally, CNNs are considered black-boxes as their operation involves predicting the outcome without the knowledge of how the prediction was made [14]. This is particularly true for applications in healthcare as there should be a reliable evidence-based rationale for the diagnosis. The fact that the model operates accurately and yet it is not interpretable leads to a significant loss of credibility [13].

In this regard, Explainable Artificial Intelligence (XAI) has come to the forefront. Explainable AI techniques are capable of generating visual and numerical explanations for the prediction made by a machine learning algorithm to improve its interpretability and transparency. XAI techniques have been extensively used to localize the areas of interest from the MRI scans and understand which features of the MRI are important for the decision made by the model [9], [21]. Such a feature allows medical practitioners to determine whether or not the focus of the model was on the right area of the MRI scan while diagnosing the tumor.

Moreover, some recent research has focused on ensemble or hybrid models using different deep learning approaches along with XAI techniques to increase both the efficiency of the process and its explainability [4]. Such approaches do not only lead to increased performance but also provide multiple ways of explanation, thus making the application of the AI approach more trustworthy. However, such drawbacks as low generalizability of results across various datasets, the usage of 2D instead of 3D images, and insufficiently defined metrics for the evaluation of XAI remain.

Multiple studies have explored individual CNN architectures for brain tumor classification; limited work has systematically compared multiple state-of-the-art transfer learning models under a unified preprocessing and evaluation framework while integrating explainable AI for interpretability. This study addresses this gap through a comparative analysis of ResNet50, DenseNet121, EfficientNetB0, and InceptionV3 combined with Grad-CAM visualization. The process will involve an analysis of cutting-edge models with the application of XAI to ensure an interpretable solution. The

goal is to create a safe, quick, and clinically applicable AI diagnosis tool for identification brain tumors.

2 Literature Review

The development of MRI technology in neuro-oncology via the application of deep learning has led to the revolution of diagnosing and classifying brain tumors. The problem of brain tumors involves excessive growth of the cells in the brain. It is often challenging to diagnose these types of tumors owing to their complex composition and form. The use of MRI technology for diagnosing the brain tumors becomes essential because of its good ability to show soft tissues and its multiple contrast imaging such as T1, T2, and FLAIR images [7]. However, it is a time-consuming process, making it subjective as well [10].

Classification of brain tumors was achieved through conventional machine learning approaches such as SVM, RF, and k-NN, which required the manual process of feature selection that included techniques like intensity and texture [15]. Although these approaches showed good results, their inability to learn complex features led to poor results on out-of-sample data [5]. The rise of deep learning models, particularly CNNs, marked the onset of a new era, where learning directly from raw MRI scans became possible without any feature selection [6].

The recent advances (2019-2025) have exhibited that the deep CNN structure can provide extremely good classification performance. As an instance, residual networks, or ResNets, incorporate skip connections in order to avoid the issue of vanishing gradients, hence learning the deep network. ResNets were found capable of achieving classification accuracy of more than 97% for multi-class brain tumor classification [2]. In contrast, DenseNet models improve feature reuse with dense connections and thus learn deep structures with fewer parameters [4]. The capability of DenseNet models to capture critical fine details of the tumor has been proven important in medical field [12].

In terms of compound scaling, EfficientNet has also challenged the limits by scaling the network's depth, width, and resolution at once. The architecture is advantageous due to the balance it creates between increasing accuracy and reducing computational costs, making it well suited for the field of medical imaging. Studies have demonstrated that a model based on the EfficientNet architecture may achieve up to 99.85% accuracy in medical diagnosis applications [37]. Whereas, the Inception family utilizes convolutional layers of different dimensions in order to obtain local and global features from images, but requires larger datasets [10].

Other than CNN models, more advanced MLPs models like ViTs and ConvNeXt models that utilize more contemporary neural net architecture design methodologies have been studied, where these models use attention mechanism for learning long-range dependencies from MRI image data. The latest literature indicates that these models perform better for generalization and classification tasks and give high accuracies (> 98%) on various datasets. Ensemble models consisting of multiple variations of models have been proved more stable and accurate [4].

However, despite the success, the uninterpretable nature (black box) of such models is a challenge that needs to be addressed, more so in medical applications, whereby interpretability is important for the effective utilization of such models in making decisions. Therefore, XAI approaches have become increasingly common to enable interpretability in deep neural networks. Techniques like Grad-CAM provides visual evidence by means of heatmaps to reveal which features the model takes into account, which can then be verified by clinicians to ensure that the model analyzes relevant features associated with tumors [9]. Similarly, LIME and SHAP provide model-agnostic approaches to XAI.

In order to address obstacles like uncertainty, the use of multiple XAI tools has become imperative in the recent literature to ensure increased validity. As an example, ensemble XAI methods combining Grad-CAM, SHAP, and LIME have proved to increase interpretability and verify model predictions [1], [4]. Furthermore, it was found that XAI tools are capable not only to provide an explanation for the models' predictions but also to help in optimizing the models through removal of redundant features [1]. This trend proves the increasing importance of explainability as a factor ensuring the success of a model in terms of all three pillars of success [4].

However, there are also several limitations of the research. First, an important drawback that exists is the problem with generalization of data because some models that were built specifically for certain data sets cannot be generalized to other data sets due to different imaging protocols [4]. Second, the vast majority of methods used 2D slicing instead of recognizing the 3D structure of MRI images, thus reducing the effectiveness of diagnosis [8]. Third, there are currently no standardized criteria for evaluating the effectiveness of XAI, and thus validating its explanations becomes problematic [14]. Fourth, very few studies have addressed multiple-data source research.

3 Methodology

Illustrated in Fig. 1 displays the overall workflow of the intended research. The framework consists of data preprocessing, model training using multiple deep learning architectures, performance evaluation, and explainability analysis using Grad-CAM.

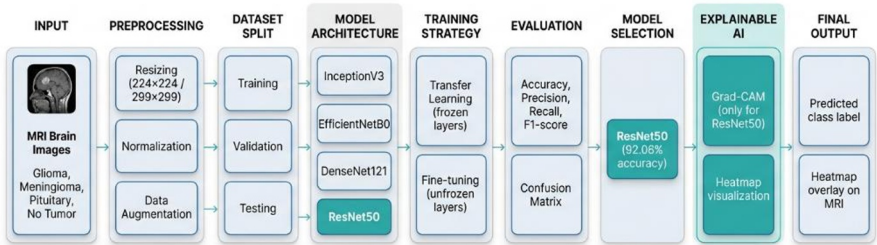


Fig. 1. Research workflow for brain tumor classification using DL models and explainable AI (Grad-CAM).

For this experiment, we have selected a data set that contains MR images of brain tumors [16]. It includes 7,200 MR images of brain tumors divided into four significant categories: glioma, meningioma, pituitary tumors, and no tumors. These images are segregated into the training and test datasets equally in terms of all classes. Specifically, the training data-set includes 1,400 images per category (total 5,600 images), while the test dataset includes 400 images per category (total 1,600 images). The data set is constructed by combining multiple publication data sets, including Figshare brain MRI data set, SARTAJ data set, and Br35H data set. The pre-processing steps include the removal of duplicate images belonging to the same category and avoiding any overlap between the training and testing sets. The dataset was divided using an 80:20 validation split from the training set.

To train deep learning models, the resolution of images was unified through the process of resizing them to a particular resolution (either 224×224 or 299×299 pixels depending on the model). Additionally, the pixel values for each image were normalized depending on the selected model. This can be summarized as follows:

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

where (x) is the original pixel value, and (μ) and (σ). Further, the input data augmentation through rotations, horizontal flips, and zooming was also employed while training for better generalization of the model and resistance against changes in tumor size and orientation.

Furthermore, the proposed methodology utilizes different kinds of CNN model architectures using the transfer learning approach, which include InceptionV3, EfficientNetB0, DenseNet121, and ResNet50 models, among others. These models are chosen based on their ability to produce outstanding results when classifying medical images. Convolutional neural networks rely on hierarchical feature extraction through the convolution method. This mathematical formula explains the process of convolution as follows:

$$f(i, j) = \sum_m \sum_n I(i - m, j - n) \cdot K(m, n) \quad (2)$$

where (I) is the input image and (K) denotes the convolutional kernel. The model had used weights pretrained on the ImageNet dataset, and therefore features such as edges and textures could be inherited. The classification layers that came after that were changed into a fully connected head with global average pooling, dense layer, dropout, and softmax output layer. There is one model among the three models that uses residual learning to tackle the problem of vanishing gradients and therefore can be used to train deep neural networks.

Regarding the training procedure, the two-stage approach was used to ensure efficiency of the feature extraction and tuning of the networks for solving the specific problem. As for the first stage of training, it was conducted exclusively for the classifiers in the network while the rest was kept frozen to facilitate tuning the pretrained network to brain tumor classification. In the 2nd stage of fine-tuning, the upper layers of the network were tuned using a lower learning rate. Early stopping and learning rate reduction strategies applied to prevent overfitting and optimize convergence.

The models were trained applying the Adam optimizer with an initial learning rate of 1×10^{-4} during feature extraction and 1×10^{-5} during fine-tuning and categorical cross-entropy loss. A batch size of 32 was used for all experiments. Softmax activation was used in the output layer to calculate probabilities:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (3)$$

The validation data was used to observe how well the model learns, and methods like early stopping and learning rate reduction techniques were applied to avoid overfitting and obtain optimal convergence.

The performance of all models was measured according to the usual classification criteria such as accuracy, precision, recall and F1-score. The model with the best results was ResNet50, which showed accuracy of 98.09% in the training set and 92.06% in the test one. Thus, it can be stated that the model managed to find a pattern that helps to distinguish classes and generalize new cases. Therefore, the ResNet50 was chosen as the best-performing model to interpret.

In order to make the suggested model more interpretable, it was necessary to use Grad-CAM on ResNet50. Grad-CAM emphasizes important segments of the image that help to produce a certain prediction, thus being extremely useful in the medical field due to necessity of explaining the process of decision making. Grad-CAM heatmap is built on the basis of gradients of the class of interest with respect to the last convolution layer activation maps:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \quad (4)$$

where (A^k) represents the feature maps and (α_k^c) denotes the importance weights derived from gradients. The heatmap will be superimposed on the MRI image so that the contributing regions can be identified. Only the ResNet50 model was selected for the application of Grad-CAM since this is the model that showed the highest level of performance out of all the models considered. This will ensure that important information can be extracted from the best-performing model.

4 Results and Discussion

Our method was evaluated using four pre-trained DL architectures, which were ResNet50, DenseNet121, EfficientNetB0, and InceptionV3. Our method was tested on these four deep learning models by using accuracy, precision, recall, and F1-score. The result shows that there was a high performance achieved by these four pre-trained DL models in terms of classification, but with varied generalization power over various types of tumors. ResNet50 was the best model based on the highest test accuracy of 92.06%, followed by InceptionV3, whose accuracy was 91.13%, then EfficientNetB0, whose accuracy was 87.25%, and DenseNet121, with an accuracy of 86.81%.

Table 1. Performance evaluation table

Model	Accuracy (%)	Precision	Recall	F1-Score
ResNet50	92.06	0.92	0.92	0.92
InceptionV3	91.13	0.91	0.91	0.91
EfficientNetB0	87.25	0.88	0.87	0.87
DenseNet121	86.81	0.87	0.87	0.86

Tumor-class specific metrics provide insights into performance. ResNet50 is good at detecting all the tumor classes, with the perfect recall (1.00) observed for the no-tumor class and near perfect recall (0.99) for the pituitary class. However, the recall rate of the glioma class (0.80) is low compared to other classes due to difficulties in diagnosing gliomas since they have diverse morphological features that overlap with those of other tumors. The gap between training (98.09%) and testing accuracy (92.06%) indicates slight overfitting, which is common in deep learning-based medical image classification as a result of limited dataset diversity. However, the use of data augmentation, dropout, early stopping, and transfer learning helped reduce excessive overfitting and improved generalization performance. Similarly, InceptionV3 model performs almost the same as ResNet50 with perfect recall (1.00) for the pituitary and no-tumor classes and low recall for the glioma class. On the other hand, EfficientNetB0 and DenseNet121 have low recall rates for the glioma and meningioma classes, which can be attributed to their less sensitivity to the borders of tumors (See Fig.2) Based on these results, it seems that ResNet50 with deep residual learning can be superior in extracting features.

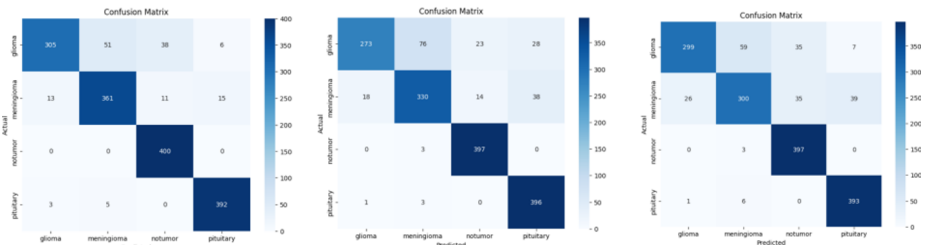


Fig. 2. Confusion matrix of (1) Inception V3, (2) Efficient net B0 and (3) Dense net 121

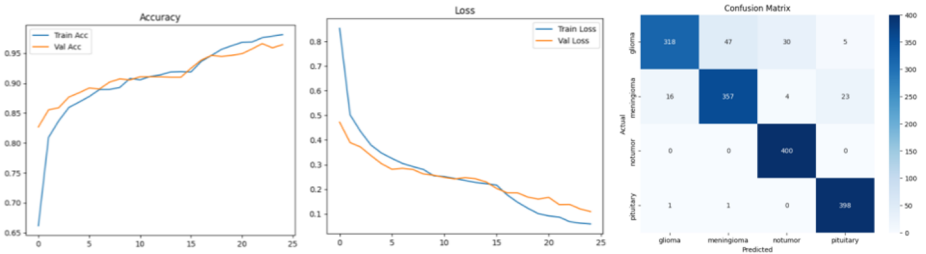


Fig. 3. Accuracy & Loss Graphs with confusion matrix for Resnet-50 model (Best Model)

All the plots indicate convergence where the training accuracy becomes higher than the validation accuracy at the end. The plot for ResNet50 indicates convergence with little variance, making it a generalized model (refer to Fig. 3). On the other hand, the plot for EfficientNetB0 and DenseNet121 reveals slight overfitting in the models since the loss in the validation set increases after several epochs. For further classification accuracy analysis, the confusion matrices of the models were evaluated (refer to Fig. 2). From the confusion matrix, it can be seen that the errors made by ResNet50 are mainly in the distinction between the glioma and meningioma classes since they have the same appearance.

To enhance the understanding of the model, the best-performing Res-Net50 model was utilized to generate heat maps using Grad-CAM. These heat maps represent the area or portion of the image that the model focuses upon. With the help of Grad-CAM images, it is evident that the model concentrates on the regions which are significant to identify the tumor, such as irregular mass regions and regions of high intensity in MRI images. This proves that the model learns no spurious correlation in the background; rather, it learns significant features for tumor identification.

Although Grad-CAM improves interpretability by highlighting tumor-relevant regions, the generated heatmaps were not clinically validated by expert radiologists, which remains a limitation of the current study. Future work should involve expert evaluation of XAI outputs to ensure clinical reliability and diagnostic relevance.

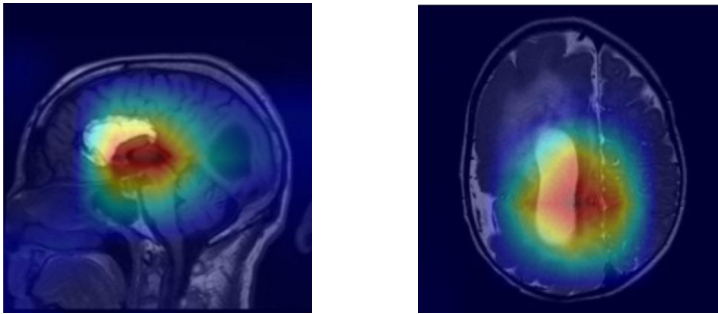


Fig. 4. Grad-CAM Output Images

By performing this comparative study, it has been proven that ResNet50 outperforms other neural networks, owing to the use of residual connections that allow deeper learning without vanishing gradients. EfficientNetB0 is efficient, but the DenseNet121 network that supports the reuse of features has marginally lower performance in recognizing tumor complexities than ResNet50. The performance of InceptionV3 is commendable but requires larger amounts of data. This research has proved that the architecture used by the neural network is essential in capturing the small differences in the medical images and the use of residual connections is vital in doing so. Further, the technique of using Explainable AI using Grad-CAM has shown how transparency can be implemented in deep learning, which was one of the biggest obstacles for health care applications of DL.

5 Conclusion

The paper introduced a comparative DL framework for multi-class brain tumor classification using MRI images by analyzing four pre-trained CNN models: InceptionV3, EfficientNetB0, DenseNet121 and ResNet50. Out of all the aforementioned models, ResNet50 proved to be most effective with an accuracy rate of 98.09% during training and 92.06% during testing due to the ability of the algorithm to learn discriminative tumor features. In addition, the usage of Explainable Artificial Intelligence (XAI) in the form of Grad-CAM provided better insight into MRI images by focusing on clinically meaningful areas. However, there are a number of limitations of this study. First of all, experiments were performed using a public dataset that is relatively small and balanced, and thus, cannot be generalized to other environments. The performance of the algorithm may decrease when working with noisy MRI images. Moreover, explainability method Grad-CAM was the only one used in this study, and its conclusions have not been verified by clinicians. In future work, we will explore the evaluation of the proposed approach on larger datasets involving multiple institutions, 3D MRI analysis, and application of other XAI methods such as SHAP and LIME. Another potential research area involves multimodal data fusion and deep learning-based hybrid models.

Acknowledgments. We, the authors would like to thank Prof. Yogita D. Patil and Vishwakarma University, Pune, for their valuable support and guidance in the present study. This research was done without any financial support from any agency.

Disclosure of Interest. We, the authors declare that there are no conflict of interests.

References

1. Iftikhar, S., Anjum, N., Siddiqui, A. B., Ur Rehman, M., & Ramzan, N. (2025). Explainable CNN for brain tumor detection and classification through XAI based key features identification. *Brain Informatics*, 12(1). <https://doi.org/10.1186/s40708-025-00257-y>
2. Singh, R., Gupta, S., Bharany, S., Almogren, A., Altameem, A., & Ur Rehman, A. (2024). Ensemble Deep Learning Models for Enhanced Brain Tumor Classification by Leveraging ResNet50 and EfficientNet-B7 on High-Resolution MRI Images. *IEEE Access*, 12, 178623–178641. <https://doi.org/10.1109/access.2024.3494232>
3. Ünlüleblebici, S., Nur Kaçmaz, R., & Öztürk, C. (2026). A Systematic Review of Machine Learning and Deep Learning Techniques for Brain Tumor Classification, and Grading. *IEEE Access*, 14, 18075–18098. <https://doi.org/10.1109/access.2026.3659641>
4. Pant, K., Dutta Pramanik, P. K., & Zhao, Z. (2026). A Robust ConvNeXt-Based Framework for Efficient, Generalizable, and Explainable Brain Tumor Classification on MRI. *Bioengineering*, 13(2), 157. <https://doi.org/10.3390/bioengineering13020157>
5. Bouhaffra, S., & El Bahi, H. (2025). Deep Learning Approaches for Brain Tumor Detection and Classification Using MRI Images (2020 to 2024): A Systematic Review. *Journal of imaging informatics in medicine*, 38(3), 1403–1433. <https://doi.org/10.1007/s10278-024-01283-8>

6. Neamah, K., Mohamed, F., Adnan, M. M., Saba, T., Bahaj, S. A., Kadhim, K. A., & Khan, A. R. (2024). Brain Tumor Classification and Detection Based DL Models: A Systematic Review. *IEEE Access*, 12, 2517–2542. <https://doi.org/10.1109/access.2023.3347545>
7. Rasheed, M., Jaffar, M. A., Akram, A., Rashid, J., Alshalali, T. A. N., Irshad, A., & Sarwar, N. (2025). Improved brain tumor classification through DenseNet121 based transfer learning. *Discover oncology*, 16(1), 1645. <https://doi.org/10.1007/s12672-025-03501-3>
8. Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R., & Bridge, C. P. (2025). A review of deep learning for brain tumor analysis in MRI. *NPJ precision oncology*, 9(1), 2. <https://doi.org/10.1038/s41698-024-00789-2>
9. Thottempudi, P., Acharya, B., Aouthu, S., Dhanalakshmi, N., Bai, B. M., Madhavi, K. R., Swaraja, K., & Malik, S. (2025). An explainable hybrid CNN–Transformer framework with Aquila optimization for MRI-based brain tumor classification. *medRxiv*. <https://doi.org/10.1101/2025.10.14.25338038>
10. Rastogi, D., Johri, P., Donelli, M., Agarwal, T., Tiwari, S., & Singh, P. (2025). XAI-BT-EdgeNet: explainable edge-aware deep learning with squeeze-and-excitation for brain tumor detection and prediction. *Frontiers in Artificial Intelligence*, 8. <https://doi.org/10.3389/frai.2025.1676524>
11. Bukaita, W., & Vadde, V. (2025). Comparative Evaluation of CNN and ResNet18 Architectures for MRI-Based Brain Tumor Classification Using Deep Learning. *Medical Research Archives*, 13(12). <https://doi.org/10.18103/mra.v13i12.7100>
12. Srinivas, V. R., & Parvathi, R. (2026). Explainable AI-driven MRI-based brain tumor classification: a novel deep learning approach. *Frontiers in Artificial Intelligence*, 8. <https://doi.org/10.3389/frai.2025.1700214>
13. Rasool, N., Wani, N. A., Bhat, J. I., Saharan, S., Sharma, V. K., Alsulami, B. S., Alsharif, H., & Lytras, M. D. (2025). CNN-TumorNet: leveraging explainability in deep learning for precise brain tumor diagnosis on MRI images. *Frontiers in oncology*, 15, 1554559. <https://doi.org/10.3389/fonc.2025.1554559>
14. Ahmed, F., Naz, N. S., Khan, S., Rehman, A. U., Ismael, W. M., & Khan, M. A. (2026). Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges. *BMC medical imaging*, 26(1), 37. <https://doi.org/10.1186/s12880-025-02118-w>
15. Biswas, A., & Islam, M. S. (2023). A Hybrid Deep CNN-SVM Approach for Brain Tumor Classification. *Journal of Information Systems Engineering and Business Intelligence*, 9(1), 1–15. <https://doi.org/10.20473/jisebi.9.1.1-15>
16. M. Nickparvar, “Brain Tumor MRI Dataset,” Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

