Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

An Explainable MRI-based Brain Tumor Diagnosis Framework with Grad-CAM Visualization

Mouthika G.[1], Hariganesh V.[2], Ganesh Srinivasan[3], Ashika Sharon L.[4], Jayanthy A.K.[5]

[1,2,3,4]Student, Department of Computer Science and Medical Engineering (Artificial Intelligence and Data Analytics), Sri Ramachandra Faculty of Engineering and Technology, SRIHER, Chennai, India.
[5]Professor-Research, Centre for Excellence, Rajalakshmi Engineering College, Chennai, India.

E-mail: [1]gmouthik27@gmail.com, [2]e0524007@sriher.edu.in, [3]e0524016@sriher.edu.in, [4]e0524025@sriher.edu.in, [5]jayanthy.ak@rajalakshmi.edu.in

Abstract

This research work presents an explainable deep learning approach to automatically detect, classify, localize, and estimate the severity of the brain tumors using MRI (Magnetic Resonance Imaging) images. It involves a VGG16-based transfer learning model with a two-stage approach of classification including first binary tumor detection and then multiclass classification of tumors. Preprocessing and data augmentation of brain MRI images is carried out before classification. Grad-CAM is applied to the proposed approach to visualize the activation map highlighting the most relevant areas of the image used by the classifier. The activation maps can be further used to approximate the localization and pixel-based severity estimation of the tumor. The experiments were conducted using a publicly available MRI dataset, and 97.56% accuracy for the binary classification and 90.38% for the multiclass classification were achieved.

Keywords: Brain Tumor Detection, Magnetic Resonance Imaging (MRI), Deep Learning, VGG16, Transfer Learning, Explainable Artificial Intelligence (XAI), Grad-CAM.

DOI: https://doi.org/10.36548/jscp.2026.3.004
Extended version of article received from 6th **International Conference on Artificial Intelligence, 6G Communications and Network Technologies 2026**
Received: 02.05.2026, received in revised form: 05.06.2026, accepted: 28.06.2026, published: 07.07.2026

1. Introduction

Brain tumors are one of the biggest challenges in the field of neurological disorders and have a major impact on morbidity and mortality worldwide. They result from uncontrolled proliferation of abnormal cells and are classified as benign or malignant. Early diagnosis is important to prevent irreversible neurological damage and to improve treatment efficacy, which may include surgery, radiotherapy or chemotherapy, thus improving survival and quality of life for patients.

MRI is the most widely used technology to evaluate brain tumors due to the excellent soft tissue contrast and spatial resolution and does not carry the risks of ionizing radiation. MRI allows visualization of intracranial structures and detection of tumor characteristics and abnormality of adjacent tissue. However, manual interpretation of MRI is labor-intensive and requires significant expertise. The diagnosis is complicated due to tumor variability, irregularities and difficulty in differentiating malignant from healthy tissues. This complexity is compounded by increasing clinical workloads that can hinder timely diagnosis and lead to inconsistencies between observers, accentuating the urgency for automated and reliable diagnostic solutions.

The recent developments in artificial intelligence, especially deep learning, have transformed the field of medical imaging, enabling computers to automatically extract complex features from unprocessed images of MRIs. CNNs have demonstrated superior capabilities in medical imaging applications through automatic extraction of complex hierarchical representations without any external help. Transfer learning takes advantage of pre-trained models and uses them in the medical domain with small training data sets.

Many classifiers focus solely on high accuracy without providing an understandable reasoning behind their decisions. Such a classifier raises questions among clinicians regarding its validity, thus making it unsuitable for use in the medical process. Besides, several classification systems only classify tumors and provide no further information on their localization or intensity. To address these issues, the proposed research involves the design of an explainable deep learning architecture dedicated to the automated identification, classification, localization, and severity assessment of brain tumors using MRI. The architecture uses the technique of transfer learning based on the pre-trained VGG16 neural network and features a two-stage classification process. The first stage includes the usage of the binary classifier to determine whether there is a brain tumor; then the image undergoes processing by the multiclass classifier that determines the type of tumor (glioma, meningioma, or pituitary tumor). In order to improve transparency of the model, Grad-CAM is used to generate heatmaps and to perform localization and severity assessment of the tumor.

2. Literature Review

Manual analysis of MRIs is very time-consuming and is largely influenced by radiologists' experience. Thus, inter-reader variability and late diagnosis are among the drawbacks of conventional techniques. In order to cope with these problems, the application of AI and deep learning becomes quite efficient for automated detection and classification of brain tumors. Specifically, Convolutional Neural Networks (CNNs) have shown tremendous potential for automatic feature extraction from MRI images without any need for feature engineering [1], [2].

Various types of CNNs and transfer learning approaches have been explored recently for multiclass brain tumor classification. The comparative evaluation of different pre-trained models like VGG16, ResNet50, InceptionV3, Xception, MobileNetV2, and EfficientNet reveals that transfer learning can be effectively employed to improve the classification performance and reduce the complexities of the training process [1]. The improvement in the architecture of the CNNs has led to better feature representations for brain tumors, and hence the discrimination between gliomas, meningiomas, pituitary, and normal brain MRIs becomes easier [3]. Furthermore, the use of hybrid deep learning approaches using multiple convolution layers and better feature representations has been found more effective for classification [4].

Transfer learning has become an essential technique for medical image analysis because of the limited availability of large annotated brain MRI datasets. Fine-tuning pretrained networks allows models to adapt effectively to medical imaging tasks while minimizing overfitting [6]. Additionally, image transformation approaches including image rotation, flipping, zooming, scaling, and brightness modification have gained widespread popularity in order to add diversity to data and increase generalizability of the model [7]. Ensemble learning approaches, which are formed using different deep learning classifiers, have also improved their robustness in terms of reducing prediction bias and ensuring accurate predictions [8]. Besides, attention based deep learning architectures have shown the capability of focusing on clinically important tumor areas [9].

Despite the remarkable success of deep learning models, their lack of interpretability remains a significant challenge for clinical implementation. Recent studies show the increasing need for XAI methods, like Grad-CAM, saliency maps, and other methods providing visual interpretation of predictions [10], [11]. These methods help clinicians to better understand the regions within images, which contribute to the classification process. Systematic reviews reveal that even though deep learning has brought significant improvements in brain tumor detection, problems of non-explainability, localization inability, insufficiency in severity assessment, and interpretability are still there [5]. Furthermore, while the methods of feature enhancement based on Histograms of Oriented Gradients (HOG) combined with deep residual networks have shown an improvement in discrimination power, they are mainly focused on classification problems, not on clinical decision making [12].

Although existing studies have achieved high classification accuracy, most focus primarily on tumor identification without providing comprehensive clinical information such as tumor localization, visual explanation, and severity estimation. Therefore, the proposed framework combines VGG16 transfer learning technique, two-step classification approach, Grad-CAM visualization interpretability, tumor location detection, and pixel-wise severity estimation all into one single architecture.

3. Methodology

This proposed framework introduces an explainable deep learning algorithm to automatically detect, classify, localize, and estimate the severity level of brain tumors based on brain MRIs. This workflow involves MRI acquisition, data set creation, image preprocessing, classification of the tumors hierarchically, explainability using Grad-CAM, tumor localization, and severity level estimation.

The architecture of the proposed framework is depicted in Figure 1. MRI images are initially preprocessed and normalized before being classified using a VGG16 transfer learning model. Initially, the proposed framework classifies whether the input MRI image has any tumor based on the input image and forwarded to a multi-class classifier to classify the image as glioma, meningioma or pituitary tumor. Finally, the result obtained is explained using Grad-CAM, to localize and estimate the severity of the tumor.

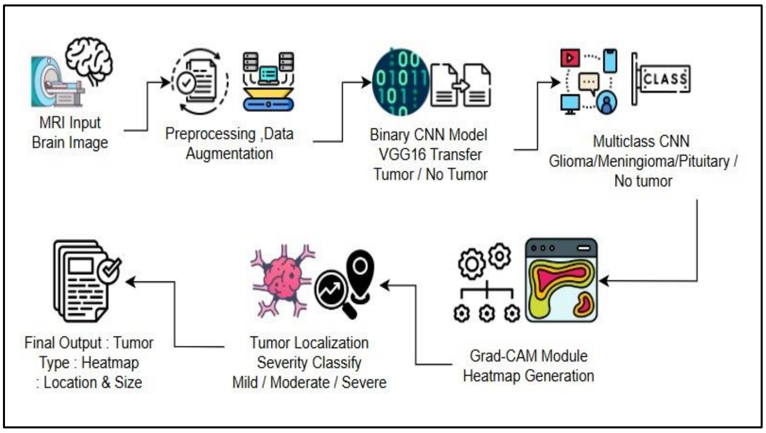

Figure 1. Proposed explainable deep learning framework for brain tumor detection and classification

In the proposed framework, a customized Brain MRI tumor classification dataset is used, which has been formed using MRI images extracted from the Figshare Brain MRI dataset [13] and SARTAJ Brain Tumor dataset [14]. Images from both datasets were combined, validated, and then reorganized to form a single dataset for multiclass brain tumor classification. This dataset consists of 7,200 images belonging to four different categories: glioma, meningioma, pituitary tumor, and no tumor. The dataset was further split into 5,600 images for training (77.8%) and 1,600 images for testing (22.2%).

Before model training, all MRI images were sorted based on their class label and resized to 224 × 224 pixels to satisfy the input constraints of the VGG16 model. Normalization of pixel values was done to enhance the convergence during the optimization process. Contrast enhancement was carried out to increase visualization of the tumors. Data augmentation such as random rotations (±15°), zooming (10%), horizontal flipping, and changing brightness (0.8–1.2) was only applied to the training set.

The resized MRI image is represented as

$$I \in \mathbb{R}^{224\times224\times3}$$

where I denotes the normalized RGB image supplied to the network. Deep feature representations are extracted using the pretrained VGG16

$$F = f_\theta(I)$$

where f_θ denotes the feature extraction function of the pretrained VGG16 model and F represents the learned feature maps.

During the first stage, the extracted features are used for binary tumor detection. The probability that an MRI image contains a tumor is computed using the sigmoid activation function

$$P_b = \sigma(W_b F + b_b)$$

where W_b and b_b denote the trainable weights and bias of the binary classifier. The network parameters are optimized using Binary Cross-Entropy loss

$$L_{\text{binary}} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i)\log(\mathbf{1} - p_i)].$$

Tumor-positive images are subsequently processed by the multiclass classifier, which predicts one of four diagnostic categories using the SoftMax activation function

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^{4} e^{z_j}}$$

where P_i denotes the probability associated with each tumor class. The multiclass classifier minimizes the categorical cross-entropy loss

$$L_{multi} = -\sum_{i=1}^{4} y_i \log(P_i).$$

The overall objective function combines both classification stages

$$L = L_{\text{binary}} + L_{\text{multi}}.$$

Network optimization is performed using the Adam optimizer

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon},$$

where η is the learning rate and $\hat{m}_t$ and $\hat{v}_t$ represent the bias-corrected first and second-order moment estimates. To preserve generic visual representations learned during ImageNet pretraining, the initial convolutional layers remain frozen, whereas approximately 3.5% of the deeper convolutional parameters are fine-tuned using the proposed MRI dataset. This transfer learning strategy enables adaptation to medical image characteristics while maintaining computational efficiency.

ISSN: 2582-2640

To improve interpretability, the proposed framework integrates Gradient-weighted Class Activation Mapping (Grad-CAM). The importance weight associated with feature map k for class c is computed as

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

where A^k denotes the activation map of the final convolutional layer.

The Grad-CAM localization map is then obtained as

$$L_{GradCAM} = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right).$$

Grad-CAM based saliency maps illustrating glioma, pituitary tumor, meningioma, and healthy brain MRI images have been provided in Figures. 2-5 respectively to prove that the proposed framework effectively identifies clinically significant tumors.

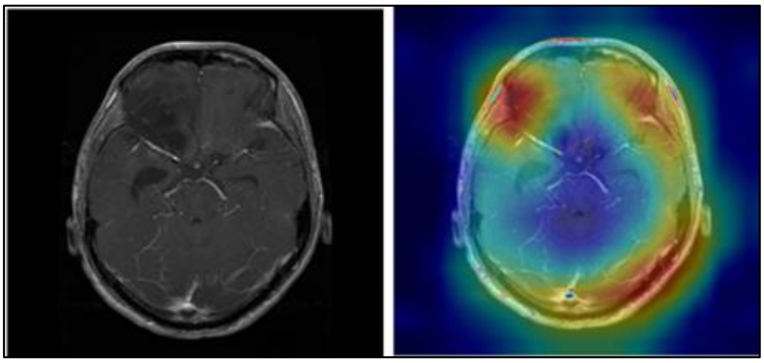

Figure 2. Grad-CAM visualization of glioma detection

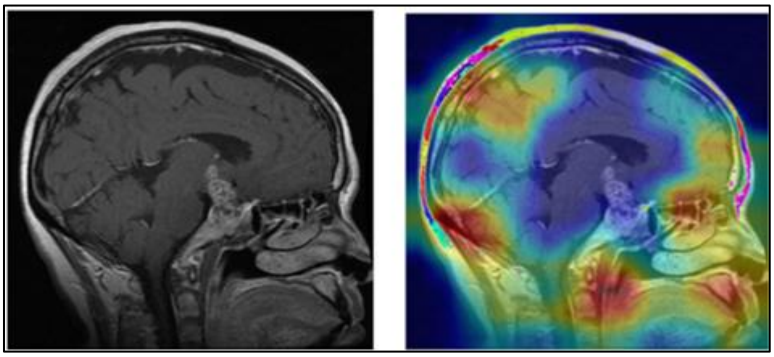

Figure 3. Grad-CAM visualization of pituitary tumor detection

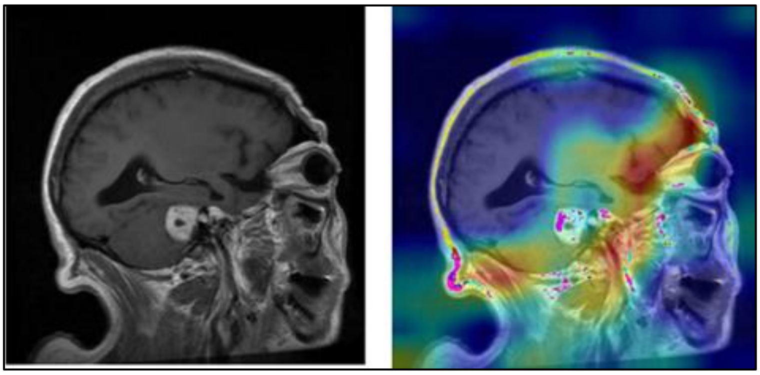

Figure 4. Grad-CAM visualization of meningioma detection

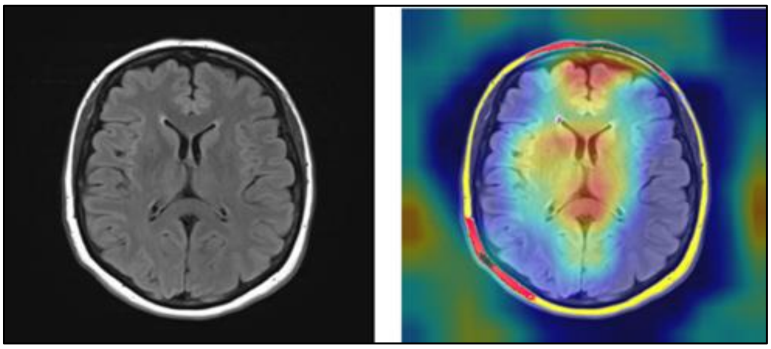

Figure 5. Grad-CAM visualization of a normal brain

The Grad-CAM activation maps are further utilized to estimate tumor location and severity. The activated tumor area is computed as

$$A_{tumor} = \sum_{i,j} M(i,j),$$

where $M(i,j)$ is a binary mask obtained by thresholding the Grad-CAM activation map.

Depending on the activation area, tumors fall into categories ranging from mild, moderate to severe based on certain cut-off criteria, thereby adding further medical information to image categorization. Thus, the entire method is a combination of proper hierarchical classification along with visualization, localization, and severity estimation, creating a complete computer-aided decision-support system.

4. Results and Discussion

Implementation of the proposed explainable deep learning model was carried out using Python 3.11 programming language with help of TensorFlow/Keras deep learning library. Training and testing were conducted on a PC with Intel Core i7 CPU, 16 GB of memory and a graphic processing unit that allows running applications using CUDA technology. The operating system was Windows 11.

The set of hyperparameters utilized for tuning the proposed VGG16-based model is shown in Table 1.

Table 1. Hyperparameter configuration of the proposed system

Hyperparameter	Value
Batch Size	32
Number of Epochs	20
Optimizer	Adam
Initial Learning Rate	0.0001

Experimental testing of the developed system was carried out by means of the independent test set of 1,600 MRI images of the customized Brain tumor classification dataset. The experiment involved such aspects as the binary classification of tumors, multiclass brain tumor classification, training convergence, and the analysis of interpretability of the prediction results. The experimental results show that the developed approach demonstrates efficient classification accuracy and stable learning properties during the training process.

The quantitative performance metrics are summarized in Table 2.

Table 2. Performance metrics of the binary brain tumor classifier

Performance Metric	Value (%)
Accuracy	97.56
Precision	99.32
Recall (Sensitivity)	97.42
Specificity	98
F1-Score	98.36

The confusion matrix for the binary classifier is shown in Figure 6. It should be noted that the classifier managed to detect 1,169 images with tumors and 392 images with normal brains. The number of false negative and false positive predictions is 31 and 8 correspondingly, which provides for 97.56% of overall classification accuracy.

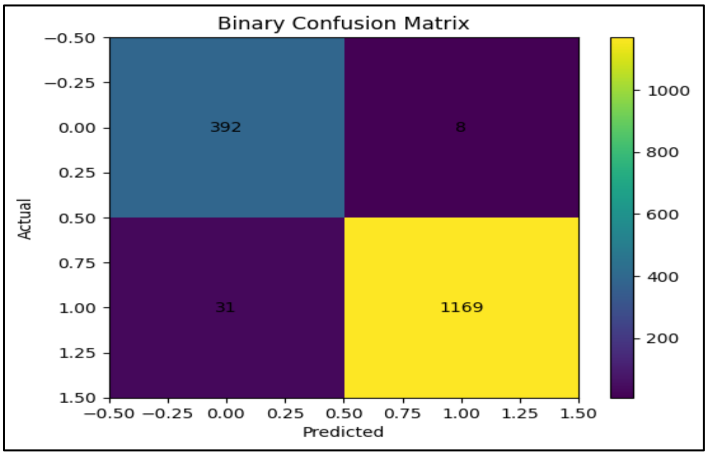

Figure 6. Confusion matrix for binary tumor classification

Class-wise evaluation metrics are presented in Table 3.

Table 3. Class-wise performance of the multiclass brain tumor classifier

Class	Precision (%)	Recall (%)
Glioma	97.61	71.75
Meningioma	82.66	91.75
No Tumor	91.9	99.25
Pituitary	92.72	98.75
Overall Accuracy = 90.38		

Multiclass classification accuracy is demonstrated in Figure 7. This classifier was tested by means of four different MRI categories: glioma, meningioma, pituitary tumor, and no tumor. 1,446 of 1,600 test images were correctly classified by the classifier, which corresponds to 90.38% of overall classification accuracy.

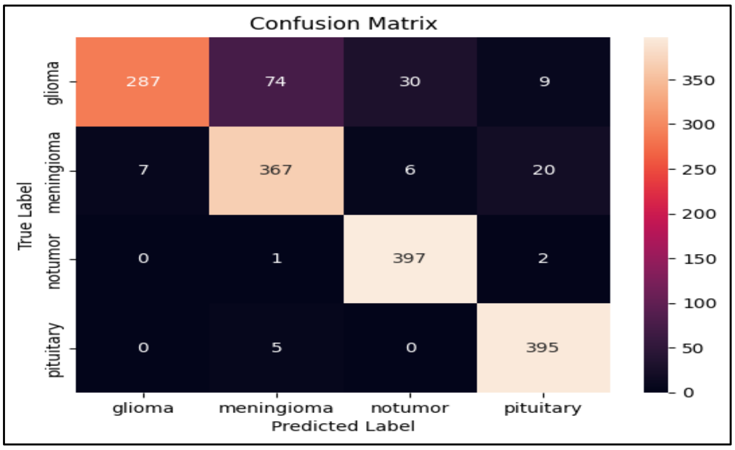

Figure 7. Confusion matrix for multiclass tumor classification

The learning process of the proposed framework is visualized through Figure 8 and 9. In particular, the binary classification model showed fast convergence with almost 99% training accuracy and 97.5% validation accuracy throughout the training process. Also, the loss curve was steadily reduced in the training process showing stable learning of the model without any instability.

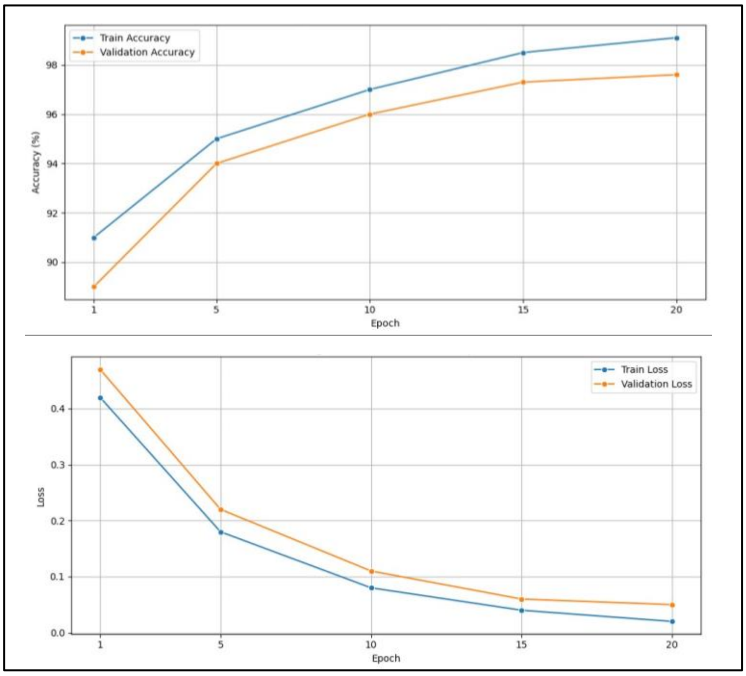

Figure 8. Binary model training and validation accuracy and loss curves

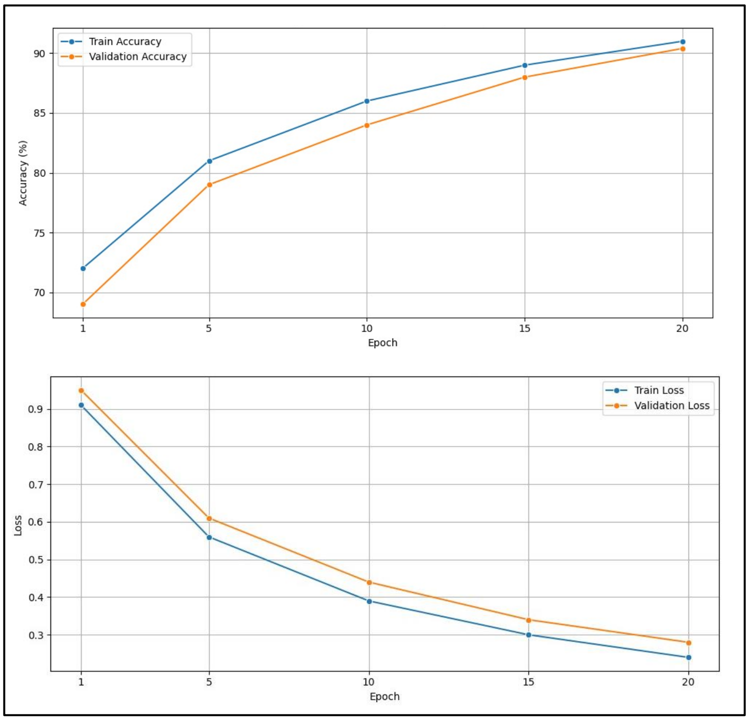

Figure 9. Multiclass model accuracy and loss across training epochs

Similarly, in the case of multiclass classification model demonstrated consistent improvement across training epochs. As shown in Fig. 9, classification accuracy gradually increased while the training loss continuously decreased, indicating successful adaptation of the fine-tuned VGG16 network to the MRI dataset

Overall, the obtained results demonstrate that the proposed framework achieves reliable classification performance while maintaining stable convergence during training. The combination of high binary detection accuracy and robust multiclass classification performance indicates that the proposed framework can effectively support automated brain tumor diagnosis and serve as a reliable computer-aided decision-support tool for MRI image analysis.

5. Conclusion

The present study offers an explainable deep learning framework for automatic brain tumor detection, classification, localization, and severity estimation through MRI image data. A transfer learning model built upon VGG16 architecture was used for binary tumor detection with following multiclass classification, which allows identifying various classes of brain tumors. Based on the experimental evaluation results, the proposed framework shows 97.56% accuracy in case of binary classification and 90.38% accuracy in case of multiclass classification, which demonstrates effective learning process on testing data set. The integration of Grad-CAM provided visual explanations that enhanced the interpretability of the model predictions and supported approximate tumor localization. Furthermore, analysis of the activated tumor regions at the pixel level can be used for severity estimation, thus providing additional clinical information apart from classification. Future work will focus on the integration of advanced segmentation networks, verification of the proposed framework on large MRI datasets across multiple institutes and clinical applications for better accuracy and efficiency.

References

[1] Gómez-Guzmán, Marco Antonio, Laura Jiménez-Beristaín, Enrique Efren García-Guerrero, Oscar Roberto López-Bonilla, Ulises Jesús Tamayo-Perez, José Jaime Esqueda-Elizondo, Kenia Palomino-Vizcaino, and Everardo Inzunza-González. "Classifying Brain Tumors on Magnetic Resonance Imaging by Using Convolutional Neural Networks." Electronics 2023, vol. 12, no. 4, 955.

[2] Wong, Yuki, Eileen Lee Ming Su, Che Fai Yeong, William Holderbaum, and Chenguang Yang. "Brain Tumor Classification Using MRI Images and Deep Learning Techniques." PloS one 2025, vol. 20, no. 5, e0322624.

[3] Aamir, Muhammad, Abdallah Namoun, Sehrish Munir, Nasser Aljohani, Meshari Huwaytim Alanazi, Yaser Alsahafi, and Faris Alotibi. "Brain Tumor Detection and Classification Using an Optimized Convolutional Neural Network." Diagnostics 2024, vol. 14, no. 16, 1714.

[4] Taher, Fatma, Mohamed R. Shoaib, Heba M. Emara, Khaled M. Abdelwahab, Fathi E. Abd El-Samie, and Mohammad T. Haweel. "Efficient Framework for Brain Tumor Detection Using Different Deep Learning Techniques." Frontiers in Public Health 2022, vol. 10, 959667.

[5] Raghavendra, U., Anjan Gudigar, Aritra Paul, T. S. Goutham, Mahesh Anil Inamdar, Ajay Hegde, Aruna Devi et al. "Brain Tumor Detection and Screening Using Artificial Intelligence Techniques: Current Trends and Future Perspectives." Computers in Biology and Medicine 2023, vol. 163, 107063.

[6] Dhakshnamurthy, Vinod Kumar, Murali Govindan, Kannan Sreerangan, Manikanda Devarajan Nagarajan, and Abhijith Thomas. "Brain Tumor Detection and Classification Using Transfer Learning Models." Engineering Proceedings 2024, vol. 62, no. 1, 1.

[7] Anaya-Isaza, Andrés, and Leonel Mera-Jiménez. "Data Augmentation and Transfer Learning for Brain Tumor Detection in Magnetic Resonance Imaging." IEEE access 2022, vol. 10, 23217-23233.

[8] Islam, Md Naim, Md Shafiul Azam, Md Samiul Islam, Muntasir Hasan Kanchan, AHM Shahariar Parvez, and Md Monirul Islam. "An Improved Deep Learning-Based Hybrid Model with Ensemble Techniques for Brain Tumor Detection from MRI Image." Informatics in Medicine Unlocked 2024, vol. 47, 101483.

[9] Banerjee, Tathagat, Prachi Chhabra, Manoj Kumar, Abhay Kumar, Kumar Abhishek, and Mohd Asif Shah. "Pyramidal Attention-Based T Network for Brain Tumor Classification: A Comprehensive Analysis of Transfer Learning Approaches for Clinically Reliable and Reliable AI Hybrid Approaches." Scientific Reports 2025, vol. 15, no. 1, 28669.

[10] Dorfner, Felix J., Jay B. Patel, Jayashree Kalpathy-Cramer, Elizabeth R. Gerstner, and Christopher P. Bridge. "A Review of Deep Learning for Brain Tumor Analysis in MRI." NPJ Precision Oncology 2025, vol. 9, no. 1, 2.

[11] Mittal, Praveen. "From Black Box AI to XAI in Neuro-Oncology: A Survey on MRI-Based Tumor Detection." Discover Artificial Intelligence 2025, vol. 5, no. 1, 1.

[12] Sharma, Arpit Kumar, Amita Nandal, Arvind Dhaka, Kemal Polat, Raghad Alwadie, Fayadh Alenezi, and Adi Alhudhaif. "HOG Transformation Based Feature Extraction Framework in Modified Resnet50 Model for Brain Tumor Detection." Biomedical Signal Processing and Control 2023, vol. 84, 104737.

[13] Cheng, Jun (2017). "Brain Tumor Dataset. " [Dataset] figshare. [Online] Available: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

[14] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. "Brain Tumor Classification (MRI) " 2025 [Dataset] Kaggle. [Online] Available: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri