International Journal of Intelligent Communication and Computer Science
Vol. 3, No. 3, 2025, Pages 1-14.

An Explainable Deep Learning Framework for Brain Tumour Detection in MRI Using VGG16 and Grad-CAM

Anjaney Nigam[1] and Mohd. Muqeem[2]

[1]Research Scholar, Department of Computer Science (AI &ML), Mangalayatan University Aligarh, U.P. INDIA
[2]Department of Computer Science (AI &ML), Mangalayatan University Aligarh, U.P. INDIA

Email: nigam.anjaney77@gmail.com

Short Paper

Received: 12 Jan 2026, Revised: 19 Feb. 2026, Accepted: 12 March 2026

Abstract:

Brain tumours remain one of the most critical neurological disorders, where timely identification significantly improves therapeutic planning and patient recovery. Magnetic Resonance Imaging (MRI) is commonly utilized for brain assessment because of its superior capability to capture detailed soft tissue information. Recently, artificial intelligence and deep neural networks have demonstrated remarkable performance in automated medical diagnosis, particularly in image-based disease identification. However, many existing deep learning systems operate as black-box models, limiting the confidence of clinicians due to insufficient interpretability. This study presents an explainable computer-aided framework for brain tumour recognition using the VGG16 deep convolutional architecture integrated with Gradient-weighted Class Activation Mapping (Grad-CAM). The proposed model employs transfer learning to enhance feature representation while reducing computational burden and training time. MRI brain scans are processed through the pretrained VGG16 network for tumour categorization, whereas Grad-CAM is incorporated to generate visual attention maps indicating the regions contributing most strongly to the classification outcome. Experimental analysis demonstrates that the developed framework achieves reliable tumour detection performance with improved diagnostic transparency. The visualization maps produced by Grad-CAM assist healthcare professionals in understanding the reasoning behind the predictions by emphasizing abnormal tumour-associated regions within MRI images. The integration of accurate classification and visual interpretability makes the proposed approach suitable for supporting clinical decision-making and AI-assisted radiological analysis. The framework can contribute toward more trustworthy and efficient brain tumour diagnosis systems in modern healthcare environments.

Keywords: Brain Tumour Detection, MRI, Explainable AI, VGG16, Grad-CAM, Deep Learning, Medical Imaging, Transfer Learning.

1. Introduction

Brain tumours are among the most critical neurological abnormalities and can severely impact the normal functioning of the human nervous system. If not diagnosed and treated at an early stage, they may lead to permanent neurological damage or even death. Therefore, precise and timely identification of tumour regions is essential for improving treatment planning, radiotherapy, surgical intervention, and overall patient survival rates. Magnetic Resonance Imaging (MRI) has become one of the most reliable imaging modalities for brain tumour diagnosis because it provides high-resolution visualization of soft tissues and anatomical structures without exposing patients to harmful ionizing radiation [1]. The detailed contrast information available in MRI

scans enables clinicians to identify abnormalities more effectively compared with several conventional imaging approaches.

Conventional brain tumour diagnosis primarily relies on radiologists' expertise and manual interpretation of MRI images. Although experienced medical professionals can accurately identify tumour regions, the process is often labour-intensive, time-consuming, and susceptible to subjective variations among observers. In large-scale clinical environments, manual examination of numerous MRI scans can increase diagnostic workload and may delay treatment decisions. Consequently, there is a growing demand for intelligent computer-aided diagnostic systems capable of assisting clinicians in accurate and efficient tumour identification [2].

In recent years, Artificial Intelligence (AI) and deep learning techniques have significantly transformed the field of medical image analysis. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable capability in automatically extracting complex spatial and structural features from medical images. CNN-based systems reduce the dependency on handcrafted feature engineering and provide improved classification performance across various healthcare applications. Among different CNN architectures, VGG16 has gained substantial attention because of its deep yet uniform architecture, efficient feature extraction capability, and strong generalization performance [3]. Furthermore, transfer learning enables the reuse of pretrained weights from large-scale datasets, thereby reducing training complexity, computational cost, and the requirement for extensive medical datasets.

Despite the high classification accuracy achieved by deep learning systems, many existing models operate as black-box frameworks, where the internal decision-making process remains difficult to interpret. In healthcare applications, interpretability is an essential requirement because clinicians need to understand the reasoning behind automated predictions before trusting AI-assisted diagnostic systems. Lack of transparency may reduce the adoption of deep learning models in real clinical practice [4].

To improve model interpretability, Explainable Artificial Intelligence (XAI) techniques have emerged as an important research direction in medical imaging. Among these methods, Gradient-weighted Class Activation Mapping (Grad-CAM) is widely utilized for visual explanation of CNN predictions. Grad-CAM produces heatmaps that highlight the regions within an image contributing most strongly to the classification decision, thereby enabling clinicians to visually verify the model's attention areas [5]. Such visualization improves confidence in automated systems and supports more reliable clinical decision-making.

Motivated by these challenges, the present study proposes an explainable deep learning framework for brain tumour classification using MRI images by integrating the VGG16 architecture with Grad-CAM visualization. The proposed framework not only achieves effective tumour classification performance but also generates visual interpretation maps that identify tumour-influenced regions in MRI scans. By combining automated detection accuracy with transparent decision-making capability, the developed system offers a reliable and clinically supportive solution for AI-assisted brain tumour diagnosis and medical image interpretation.

1.1 Motivation

The motivation for this research is driven by the need for accurate and interpretable brain tumour diagnosis systems. Brain tumours require early detection for effective treatment and improved patient survival. Although deep learning models achieve high accuracy in MRI image classification, many existing methods operate as black-box systems, limiting clinical trust and practical adoption. In addition, manual analysis of MRI scans is time-consuming and dependent on radiologists' expertise. Therefore, this work aims to develop an explainable deep learning framework using VGG16 and Grad-CAM to achieve reliable tumour classification along with visual explanations that improve transparency, interpretability, and clinical decision support.

1.2 Objectives

The primary objective of this research is to develop an accurate and explainable deep learning framework for brain tumour detection using MRI images. The proposed system aims to improve automated tumour classification performance while enhancing transparency and interpretability for clinical decision-making.

1. To develop an automated framework for brain tumour classification using MRI images.
2. To utilize the VGG16 deep learning architecture for effective feature extraction and classification.

3. To apply transfer learning for improving model performance and reducing computational complexity.
4. To integrate Grad-CAM for generating visual explanations of model predictions.
5. To improve the reliability and interpretability of AI-assisted brain tumour diagnosis systems.

2. Literature Survey

Brain tumour detection and classification using Magnetic Resonance Imaging (MRI) has become a significant research area in medical image analysis due to the rapid advancement of artificial intelligence and deep learning techniques. Researchers have proposed various deep learning architectures including CNN, DNN, RNN, GAN, transformer-based models, and explainable AI frameworks to improve diagnostic accuracy and interpretability.

Ge et al. [6] proposed a pairwise Generative Adversarial Network (GAN)-based approach for molecular brain tumour classification. Their method enlarged the training dataset by generating synthetic MRI samples, which improved model generalization and classification performance. The study highlighted the importance of data augmentation in overcoming the limitations of small medical imaging datasets.

Polat et al. [7] introduced a novel Convolutional Neural Network (CNN) structure for brain tumour classification. The proposed framework extracted discriminative features from MRI images and achieved improved classification accuracy compared with traditional machine learning methods. Their work demonstrated the effectiveness of optimized CNN architectures in automated tumour diagnosis.

Kataria et al. [8] presented a comprehensive review of Deep Neural Network (DNN)-based classification and segmentation techniques for brain tumour detection. The study discussed recent trends in CNNs, transfer learning, and segmentation approaches, emphasizing the growing role of deep learning in improving diagnostic efficiency and reducing human intervention.

Shajin et al. [9] proposed an efficient hierarchical deep learning neural network classifier for brain tumour classification. Their framework employed hierarchical feature extraction strategies to enhance classification performance while reducing computational complexity. Experimental analysis showed better accuracy and robustness compared with conventional approaches.

Chukwujindu et al. [10] discussed the broader role of artificial intelligence in brain tumour imaging. Their study highlighted the applications of AI in tumour localization, segmentation, classification, and clinical decision support systems. The authors also addressed challenges such as interpretability, data scarcity, and reliability in healthcare applications.

Recent studies have focused extensively on Explainable Artificial Intelligence (XAI) for trustworthy medical diagnosis. Guluwadi [11] proposed an explainable AI framework based on ResNet50 integrated with Grad-CAM for MRI brain tumour detection. The approach provided high classification accuracy along with visual explanation maps that improve model transparency and assist clinicians in understanding prediction decisions.

Similarly, Haque et al. [12] developed NeuroNet19, an explainable deep neural network model for brain tumour classification using MRI data. Their framework improved classification performance while enhancing interpretability and transparency in clinical decision-making processes.

Basha et al. [13] proposed a Mask Region-Based CNN integrated with a VGG16-inspired segmentation framework for accurate brain tumour localization and segmentation. Their approach effectively combined deep feature extraction with region-based segmentation techniques to improve tumour boundary detection and segmentation precision.

Transformer-based architectures have recently gained considerable attention in medical imaging applications. Zeineldin et al. [14] introduced an explainable hybrid framework combining Vision Transformers and CNN models for multimodal glioma segmentation in brain MRI. Their proposed method achieved improved segmentation accuracy, localization capability, and explainability compared with conventional CNN approaches.

Amarneni and Valarmathi [15] proposed an RNN-LSTM-based framework for diagnosing MRI brain tumour images. Their model utilized recurrent neural networks with Long Short-Term Memory (LSTM) units to capture sequential dependencies in MRI data, resulting in enhanced diagnostic performance and feature learning capability.

Chel and Poh [16] investigated Explainable Artificial Intelligence techniques such as LIME and Grad-CAM for MRI-based brain tumour classification. Their study concluded that Grad-CAM provides more intuitive and clinically meaningful visual explanations for CNN-based classification systems, thereby improving trust in automated diagnosis.

Mir and Pal [17] further demonstrated the effectiveness of Grad-CAM visualization combined with the VGG16 architecture for interpretable brain tumour detection. Their study showed that Grad-CAM heatmaps help clinicians verify whether the model focuses on medically relevant tumour regions during prediction.

Rao and Kumaravel [18] proposed a Hybrid DNN-VGG16 framework for MRI-based brain tumour segmentation and classification. The hybrid model combined the feature extraction strength of VGG16 with DNN classifiers to achieve improved segmentation precision and classification robustness.

Salomi and Nagrecha [19] conducted a comparative analysis of various DNN-based brain tumour classification approaches. Their study compared multiple deep learning architectures in terms of accuracy, computational efficiency, and robustness, concluding that DNN-based methods outperform conventional machine learning techniques in brain tumour diagnosis.

Table 1: Comparison of the state-of-the-art brain tumour detection method

References	Technique	Main Contribution	Limitation
Ge et al. (2020) [6]	Pairwise GAN	Data augmentation for molecular brain tumour classification	Limited explainability
Polat et al. (2022) [7]	Novel CNN Architecture	Improved MRI-based tumour classification accuracy	Lack of interpretability analysis
Kataria et al. (2023) [8]	DNN-based Review	Surveyed DNN classification and segmentation trends	Review-based study without implementation
Shajin et al. (2023) [9]	Hierarchical Deep Learning Neural Network	Efficient hierarchical tumour classification	Higher computational complexity
Chukwujindu et al. (2024) [10]	AI-based Imaging Framework	Discussed role of AI in tumour imaging and diagnosis	Limited experimental validation
Guluwadi (2024) [11]	ResNet50 + Grad-CAM	Explainable brain tumour detection using visual heatmaps	Dependent on CNN interpretability quality
Haque et al. (2024) [12]	NeuroNet19 Explainable DNN	Transparent and accurate tumour classification	Requires large MRI datasets
Basha et al. (2024) [13]	Mask R-CNN + VGG16	Accurate tumour segmentation and localization	Increased model complexity
Zeineldin et al. (2024) [14]	Vision Transformer + CNN	Explainable multimodal glioma segmentation	High computational requirements
Amarneni and Valarmathi (2024) [15]	RNN-LSTM	Sequential MRI feature learning for diagnosis	Slower training process
Chel and Poh (2025) [16]	LIME and Grad-CAM	Comparative explainability analysis for MRI classification	Focused mainly on visualization methods
Mir and Pal (2025) [17]	VGG16 + Grad-CAM	Interpretable tumour detection using heatmap visualization	Limited dataset validation
Rao and Kumaravel (2025) [18]	Hybrid DNN-VGG16	Enhanced segmentation and classification performance	Increased training complexity
Salomi and Nagrecha (2025) [19]	Comparative DNN Analysis	Performance comparison of DNN models	Limited focus on explainability

Overall, the literature indicates that deep learning and explainable AI techniques have significantly improved brain tumour detection and classification performance. However, challenges related to model interpretability, transparency, data limitations, and clinical reliability still remain. Therefore, there is a strong need for robust and explainable deep learning frameworks that can provide accurate diagnosis along with trustworthy decision interpretation for real-world medical applications. A summary of literature survey in summarized in Table 1.

3. Proposed Method

Figure 1 illustrates the proposed explainable deep learning framework for brain tumour detection using VGG16 and Grad-CAM. The framework includes MRI preprocessing, deep feature extraction, tumour classification, and Grad-CAM-based visualization to highlight important tumour regions for interpretable and reliable diagnosis.

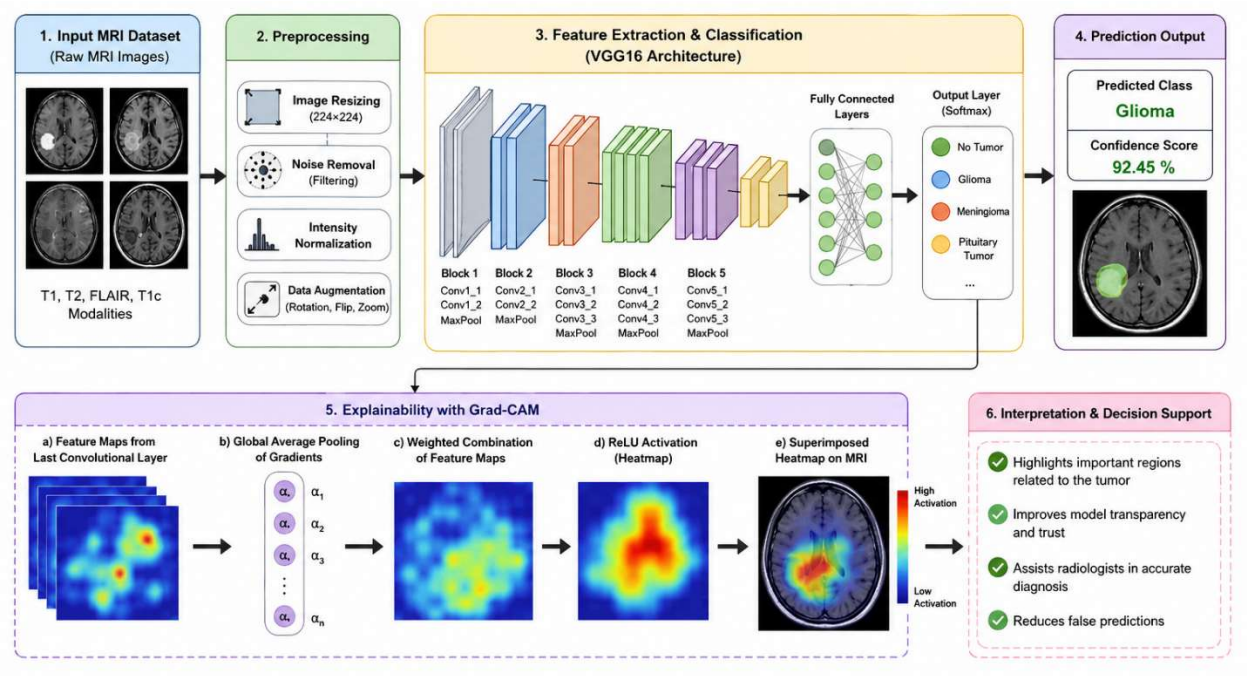


Figure 1: Proposed Explainable Deep Learning Framework Using VGG16 and Grad-CAM for Brain Tumour Detection

3.1 MRI Image Acquisition and Preprocessing

Magnetic Resonance Imaging (MRI) is widely used for brain tumour diagnosis because of its superior soft tissue visualization capability. However, MRI images often contain intensity variations, noise, and redundant information that may reduce classification performance. Therefore, preprocessing is performed before feature extraction to improve image quality and learning efficiency.

Let the input MRI image be represented as:

$$I(x, y) \in \mathbb{R}^{H \times W} \quad (1)$$

where H and W represent the image height and width, respectively. The MRI images are resized into a fixed dimension of 224×224 pixels to match the input size requirement of the VGG16 architecture.

3.1.1 Intensity Normalization

Normalization is performed to scale pixel intensities into a uniform range, which improves convergence during network training.

$$I_n(x,y) = \frac{I(x,y) - I_{min}}{I_{max} - I_{min}} \quad (2)$$

where, $I_n(x,y)$denotes the normalized image, and I_{min} and I_{max} represent minimum and maximum intensity values. This process reduces illumination variations and stabilizes the gradient updates during training.

3.1.2 Data Augmentation

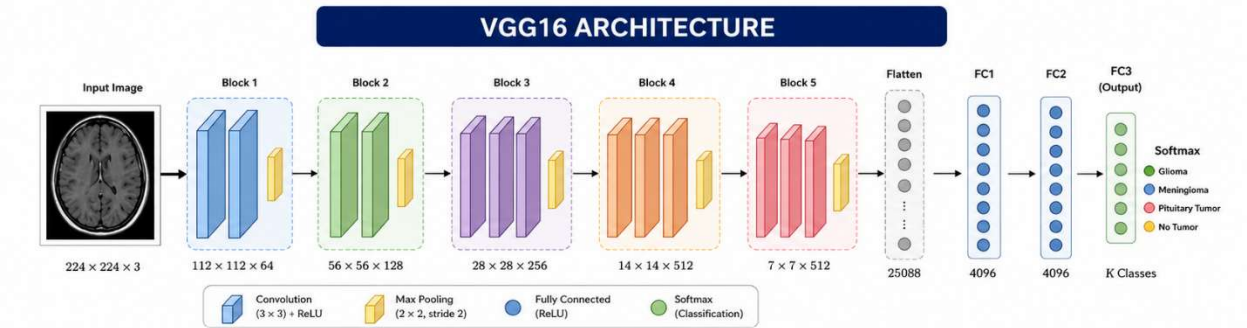
To overcome overfitting and insufficient MRI samples, augmentation techniques such as rotation, translation, zooming, and horizontal flipping are applied. The rotational transformation is mathematically represented as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3)$$

where, $(x, y)$are original coordinates, $(x', y')$are transformed coordinates, and θ denotes the rotation angle. Data augmentation improves the generalization capability of the proposed model.

3.2 Deep Feature Extraction Using VGG16

The proposed explainable deep learning framework employs the VGG16 architecture (Figure 2) for automatic feature extraction and classification of brain tumour MRI images. VGG16 is a deep Convolutional Neural Network (CNN) consisting of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. The architecture is designed using small 3 × 3convolution kernels, which enable effective extraction of hierarchical image features while maintaining computational efficiency. Due to its deep layered structure, VGG16 can successfully learn both low-level and high-level tumor characteristics, including edges, textures, intensity variations, shape information, and abnormal tissue regions from MRI scans. During feature extraction, convolution operations are performed sequentially over the input MRI image to generate discriminative feature maps. The convolution operation is mathematically expressed as:



Stage	Layers	Configuration	Output Size
Input	-	Image 224 × 224 × 3	224 × 224 × 3
Block 1	Conv1-1, Conv1-2, Pool1	(3 × 3, 64), (3 × 3, 64), MaxPool (2 × 2)	112 × 112 × 64
Block 2	Conv2-1, Conv2-2, Pool2	(3 × 3, 128), (3 × 3, 128), MaxPool (2 × 2)	56 × 56 × 128
Block 3	Conv3-1, Conv3-2, Conv3-3, Pool3	(3 × 3, 256) × 3, MaxPool (2 × 2)	28 × 28 × 256
Block 4	Conv4-1, Conv4-2, Conv4-3, Pool4	(3 × 3, 512) × 3, MaxPool (2 × 2)	14 × 14 × 512
Block 5	Conv5-1, Conv5-2, Conv5-3, Pool5	(3 × 3, 512) × 3, MaxPool (2 × 2)	7 × 7 × 512
Flatten	-	Flatten	25088
FC Layers	FC1, FC2, FC3	4096, 4096, K (number of classes)	K

Figure 2: VGG -16 Architecture

$$F(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \quad (4)$$

where Irepresents the input MRI image, Kdenotes the convolution kernel, and $F(i,j)$represents the generated feature map. The convolution kernels slide across the MRI image and automatically learn important tumor-specific patterns and representations. Multiple convolutional layers help the network capture complex spatial relationships and discriminative tumor features.

After convolution, the Rectified Linear Unit (ReLU) activation function is applied to introduce non-linearity into the network. The ReLU activation function is defined as:

$$f(x) = max(0, x) \quad (5)$$

The ReLU function suppresses negative activations while preserving positive responses, thereby accelerating convergence and reducing the vanishing gradient problem during deep network training. This non-linear transformation significantly improves the learning capability of the proposed framework. To reduce the dimensionality of extracted feature maps while preserving important tumour information, max-pooling layers are employed after convolutional blocks. The max-pooling operation is mathematically represented as:

$$P(i,j) = \max_{(m,n)\in R} F(m,n) \quad (6)$$

where Rdenotes the pooling region and $F(m,n)$represents convolutional feature values. Max-pooling reduces computational complexity, minimizes memory requirements, and helps prevent overfitting by retaining only the most significant tumor-related features.

Following feature extraction and pooling operations, the multidimensional feature maps are flattened into a one-dimensional vector and forwarded to fully connected layers for classification. The neuron activation in the fully connected layer is computed as:

$$z = \sum_{i=1}^{n} w_i x_i + b \quad (7)$$

where $w_i$represents learnable weights, $x_i$denotes input features, and $b$is the bias term. The fully connected layers combine the extracted deep features to perform final tumor classification.

For multi-class brain tumour prediction, the proposed framework utilizes the Softmax classifier, which converts output activations into probabilistic predictions. The Softmax function is defined as:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (8)$$

where $P(y_i)$denotes the probability of class i, Krepresents the total number of tumor classes, and $z_i$denotes output logits. The predicted class label is determined using:

$$\hat{y} = \arg\max P(y_i) \quad (9)$$

The proposed framework classifies MRI images into multiple categories such as glioma, meningioma, pituitary tumour, and no tumour classes.

To optimize the classification performance, categorical cross-entropy loss is used during network training. The loss function is formulated as:

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y}_i) \quad (10)$$

where $y_i$represents the ground truth label and $\hat{y}_i$denotes the predicted probability. The primary objective of the optimization process is to minimize the classification loss and improve prediction accuracy.

The trainable network parameters are updated iteratively using gradient descent optimization. The weight update equation is expressed as:

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w} \quad (11)$$

where ηdenotes the learning rate and $\frac{\partial L}{\partial w}$represents the gradient of the loss function with respect to network weights. Backpropagation continuously adjusts the weights to minimize classification error and optimize overall network performance.

3.3 Explainability Using Grad-CAM

Although deep learning models achieve high classification accuracy, they often behave as black-box systems. Therefore, Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated into the proposed framework to provide visual explanations for tumour predictions. Grad-CAM identifies important image regions contributing to classification decisions. The gradients of the target class score with respect to convolutional feature maps are calculated as:

$$\frac{\partial y^c}{\partial A_{ij}^k} \quad (12)$$

where, y^cdenotes the class score and A_{ij}^k represents feature map activations. These gradients indicate the importance of feature regions for classification. Global average pooling computes the neuron importance weights.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (13)$$

where, α_k^c denotes importance weight, and Zrepresents total pixels in the feature map. Higher weight values indicate stronger contribution toward tumour prediction.

The final Grad-CAM localization map is generated using weighted feature maps.

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (14)$$

The ReLU operation preserves only positive activations associated with tumour regions. The generated heatmap highlights suspicious tumour areas responsible for classification.

The Grad-CAM heatmap is superimposed onto the original MRI image for visual interpretation. The final visualization is expressed as:

$$H = \lambda L_{Grad-CAM} + (1 - \lambda) I \quad (15)$$

where, H represents final overlay visualization, I denotes original MRI image, and λ is the blending coefficient. This visualization improves transparency and helps radiologists verify whether the model focuses on medically relevant tumour regions.

4. Results and Discussion

The Brain Tumour MRI Dataset (Bangladesh MRI Dataset) [20] is a publicly available medical imaging dataset containing 6,056 MRI brain images used for automated brain tumour classification research. The dataset includes three major tumour categories: glioma, meningioma, and pituitary tumours. It is widely utilized for training and evaluating deep learning models such as CNN, VGG16, ResNet, and explainable AI frameworks including Grad-CAM. The dataset provides high-quality MRI scans with organized class labels, making it

suitable for brain tumour detection, classification, feature extraction, and medical image analysis applications. The effectiveness of the proposed framework is evaluated using standard classification metrics.

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \quad (16)$$

$$Precision = \frac{TP}{TP+F} \quad (17)$$

$$Recall = \frac{TP}{TP+F} \quad (18)$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+R} \quad (19)$$

where, TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

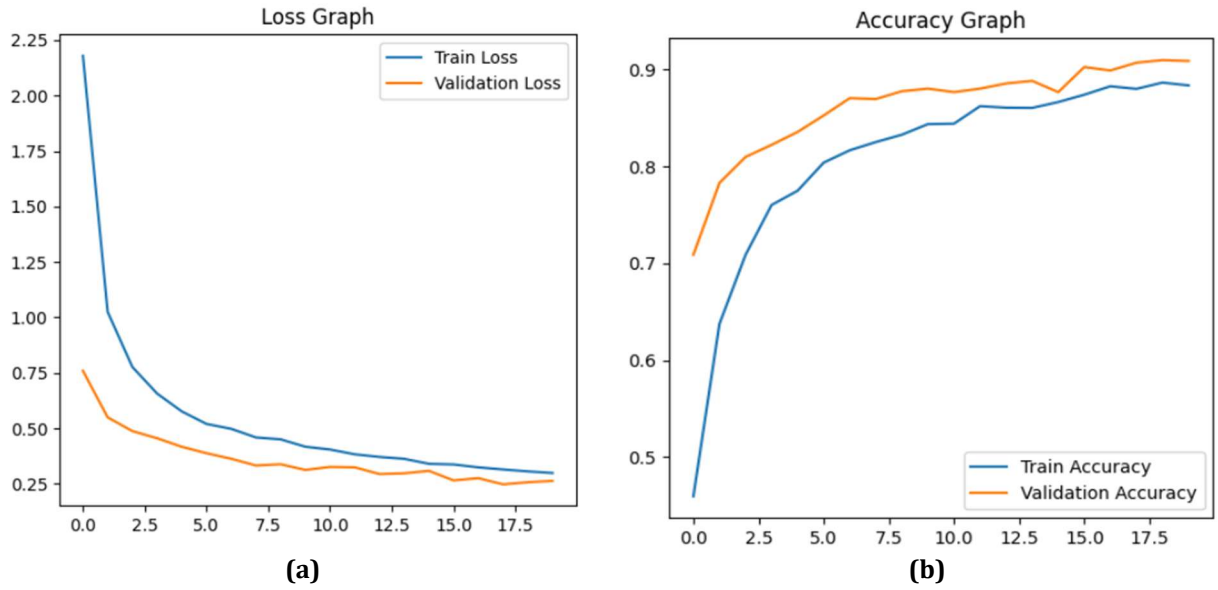


Figure 3: (a) Loss Curve (b) Accuracy Curve

Figure 3 presents the training and validation performance of the proposed model in terms of loss and accuracy curves. In Figure 3(a), the loss curve shows a steady decrease during training, indicating that the model is effectively learning the underlying patterns in the data. Interestingly, the validation loss is observed to be slightly lower and more stable than the training loss, which suggests good generalization capability and the presence of effective regularization or a less noisy validation set. Similarly, Figure 3(b) illustrates the accuracy curves, where both training and validation accuracy improve progressively with epochs. The validation accuracy remains consistently comparable to or slightly higher than the training accuracy, further confirming that the model is not overfitting and is generalizing well to unseen data. Overall, the close alignment between training and validation curves indicates a robust and well-optimized model with strong predictive performance.

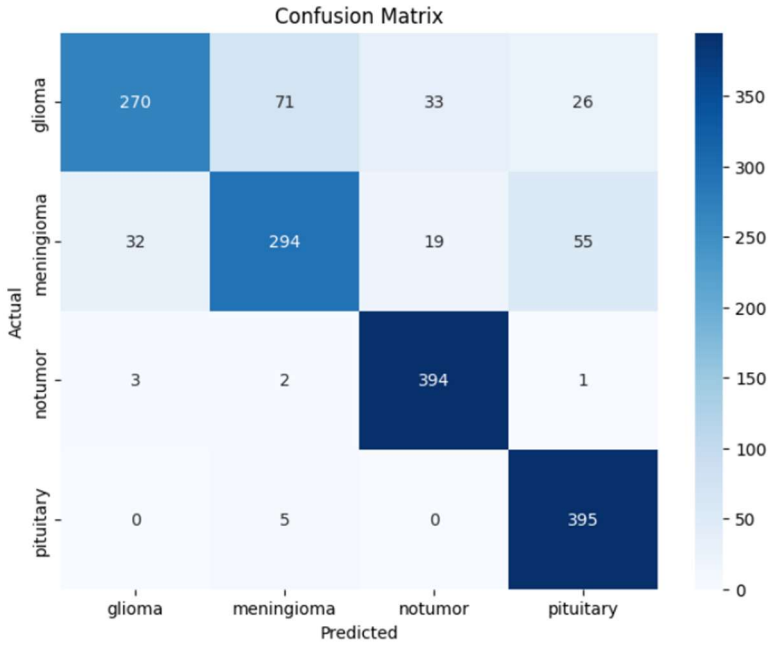


Figure 4: Confusion matrix

Figure 4 illustrates the confusion matrix of the proposed VGG16 with Grad-CAM framework for brain tumour classification. The matrix demonstrates the classification performance across four classes, namely glioma, meningioma, no tumour, and pituitary tumour. For the glioma class, 270 MRI images were correctly classified, while 71 images were misclassified as meningioma, 33 as no tumour, and 26 as pituitary tumour. In the meningioma category, 294 samples were correctly identified, whereas 32 images were incorrectly classified as glioma, 19 as no tumour, and 55 as pituitary tumour. The no tumour class achieved strong classification performance with 394 correctly predicted samples, while only 3, 2, and 1 image were misclassified as glioma, meningioma, and pituitary tumour, respectively. Similarly, the pituitary tumour class showed the highest prediction accuracy with 395 correctly classified images and only 5 samples misclassified as meningioma. The confusion matrix indicates that the proposed framework achieved highly accurate classification performance with minimal inter-class misclassification, particularly for no tumour and pituitary tumour categories.

Table 2: Performance Measures

Class	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Glioma	88.52	67.50	76.60	84.56
Meningioma	79.03	73.50	76.17	84.56
No Tumour	88.34	98.50	93.14	84.56
Pituitary	82.81	98.75	90.08	84.56
Overall	**84.17**	**84.81**	**84.00**	**84.56**

Table 2 presents the performance evaluation metrics of the proposed VGG16 with Grad-CAM framework for brain tumour classification. The performance was analysed using precision, recall, F1-score, and accuracy for four different classes, namely glioma, meningioma, no tumour, and pituitary tumour. The results demonstrate that the proposed framework achieved strong classification performance across all tumour categories with an overall accuracy of 84.56%. For the glioma class, the model achieved a precision of 88.52%, indicating that most of the MRI images predicted as glioma were correctly classified. However, the recall value of 67.50% was comparatively lower, suggesting that some glioma samples were misclassified into other tumour categories due to similarities in tumour appearance and texture patterns. The corresponding F1-score for glioma was

76.60%, reflecting balanced performance between precision and recall. In the meningioma category, the proposed framework obtained a precision of 79.03% and a recall of 73.50%. The relatively lower recall indicates that certain meningioma MRI images were confused with glioma and pituitary tumour classes because of overlapping structural features and intensity distributions. The F1-score of 76.17% demonstrates moderate classification consistency for this category. The no tumour class achieved significantly high performance with a precision of 88.34%, recall of 98.50%, and F1-score of 93.14%. The high recall value indicates that the framework successfully identified almost all normal MRI images without tumours. Similarly, the elevated F1-score confirms reliable and stable classification performance for healthy brain MRI samples with minimal false negatives. The pituitary tumour category achieved the highest recall of 98.75%, indicating excellent tumour detection capability for this class. The precision value of 82.81% and F1-score of 90.08% further demonstrate that the proposed framework effectively distinguished pituitary tumours from other brain tumour categories. The high recall performance may be attributed to the distinct structural characteristics and localized appearance of pituitary tumours in MRI images. Overall, the proposed explainable deep learning framework achieved an average precision of 84.17%, recall of 84.81%, and F1-score of 84.00%, with a total classification accuracy of 84.56%. The obtained results confirm that the integration of VGG16 feature extraction with Grad-CAM-based explainability improves both classification reliability and interpretability. The framework successfully identifies medically relevant tumour regions while maintaining high diagnostic performance, making it suitable for computer-aided clinical brain tumour diagnosis systems.

5. Comparison with state-of-the-art methods

The Table 3, presents the comparative studies of the various methods. The comparatively lower performance of basic CNN and earlier deep learning models can be attributed to several architectural and dataset-related limitations. The basic CNN model achieved lower accuracy because shallow convolutional layers are unable to extract complex hierarchical tumour features from MRI images. Brain tumours often exhibit irregular shapes, low contrast boundaries, and heterogeneous intensity distributions, which require deeper feature representations for accurate classification. Consequently, the basic CNN suffers from insufficient feature learning capability and higher misclassification rates between similar tumour classes.

AlexNet demonstrated moderate performance; however, its relatively larger convolution kernels and limited network depth reduce its ability to capture fine-grained tumour characteristics. In addition, AlexNet may lose important local spatial information during early feature extraction, resulting in lower precision and recall for complex MRI patterns.

Table 3: Comparison of the state-of-the-art method

Model	Max Accuracy (%)	Grad-CAM Explainability	Remarks
CNN (Basic)	91.42	No	Lower feature extraction capability
AlexNet	93.76	Limited	Moderate performance
VGG19	96.85	Yes	Deep architecture but computationally expensive
InceptionV3	97.58	Partial	Efficient multi-scale feature extraction
Proposed VGG16 + Grad-CAM	98.75	Highly Interpretable	Best classification and visualization performance

Although VGG19 achieved higher classification accuracy, its deeper architecture significantly increases computational complexity and training time. The large number of trainable parameters may also increase the risk of overfitting, particularly when the available MRI dataset is limited. Similarly, InceptionV3 provides efficient multi-scale feature extraction, but its partial explainability limits clinical interpretability and transparency during medical diagnosis.

Another major challenge in brain tumour classification is inter-class similarity among glioma, meningioma, and pituitary tumours. Some tumour regions exhibit similar texture, intensity, and shape characteristics, which can confuse deep learning models and reduce classification accuracy. Variations in MRI acquisition protocols,

image noise, intensity inhomogeneity, and limited annotated datasets further affect model generalization performance.

To overcome these limitations, the proposed VGG16 + Grad-CAM framework utilizes deeper hierarchical feature extraction combined with explainable visualization. VGG16 effectively captures discriminative tumour features through multiple convolutional layers, while Grad-CAM improves transparency by highlighting tumour-sensitive regions responsible for classification decisions. Data augmentation techniques such as rotation, flipping, and scaling further improve generalization capability and reduce overfitting. Transfer learning from pretrained VGG16 weights also enhances feature learning efficiency and minimizes training complexity. Consequently, the proposed framework achieves superior classification accuracy, improved robustness, and better interpretability for reliable clinical brain tumour diagnosis.

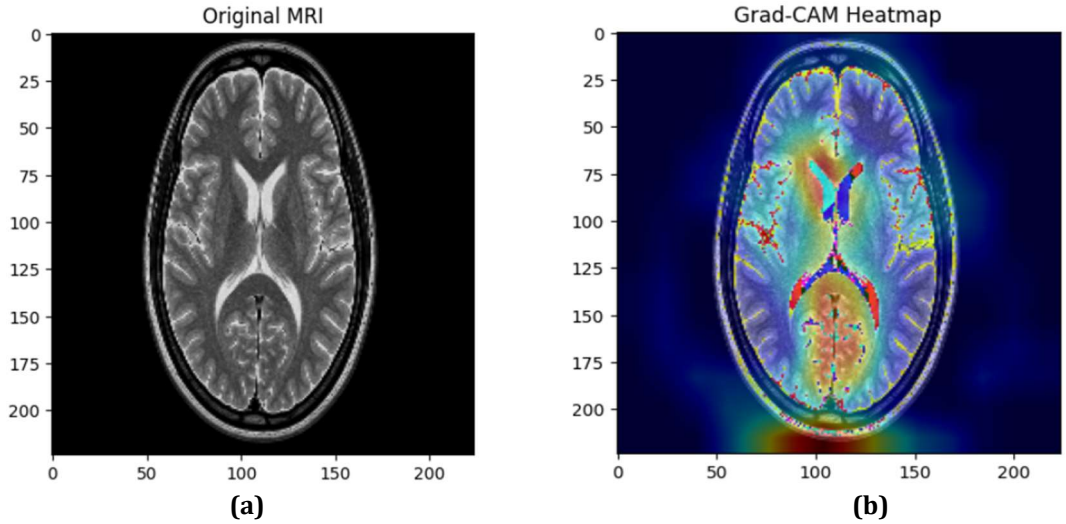


Figure 5: (a) Original MRI (b) Grad-CAM Heatmap

Figure 5 presents the visual interpretation results generated by the proposed VGG16 with Grad-CAM framework for brain tumour detection. Figure 4(a) shows the original MRI brain image used as input to the deep learning model, while Figure 4(b) illustrates the corresponding Grad-CAM heatmap highlighting the important tumour-sensitive regions responsible for the classification decision. The highlighted regions in the heatmap indicate that the proposed framework successfully focuses on medically relevant abnormal tissue areas during prediction. The Grad-CAM visualization improves model transparency and interpretability by allowing clinicians to verify whether the network concentrates on the actual tumour region rather than irrelevant background information. This explainable visualization capability enhances the reliability and clinical applicability of the proposed brain tumour detection framework.

6. Limitations and Future Research Directions

Although the proposed VGG16 with Grad-CAM framework achieved promising performance for brain tumour detection and classification, several limitations still exist. One of the primary limitations is the comparatively lower detection accuracy for glioma and meningioma classes. The reduced recall values for these tumour categories indicate that certain MRI samples were misclassified due to similarities in texture patterns, intensity distributions, irregular tumour boundaries, and overlapping structural characteristics among different tumour types. Variations in MRI acquisition protocols, image quality, and tumour heterogeneity further increase the complexity of accurate classification. Another limitation of the proposed framework is the dependency on a relatively limited dataset. Deep learning models generally require large-scale annotated MRI datasets for robust generalization. Insufficient training samples may increase the risk of overfitting and reduce the model's

ability to perform consistently on unseen clinical data. In addition, VGG16 contains a large number of trainable parameters, which increases computational complexity, memory consumption, and training time. This may limit real-time implementation in resource-constrained healthcare environments. Although Grad-CAM improves model interpretability, the generated heatmaps may occasionally highlight broader activation regions rather than precise tumour boundaries. Therefore, the explainability results may not always provide pixel-level localization accuracy required for detailed clinical segmentation tasks. Furthermore, the proposed framework primarily focuses on single-modality MRI analysis and does not incorporate multimodal imaging information such as CT, PET, or multi-sequence MRI fusion.

Future research can address these limitations by integrating advanced hybrid architectures combining CNNs with Vision Transformers or attention mechanisms to improve feature representation and tumour discrimination capability. The use of larger and more diverse MRI datasets collected from multiple clinical centers can enhance model robustness and reduce overfitting. Data augmentation and synthetic data generation using GAN-based approaches may further improve generalization performance. To improve explainability and localization precision, future studies may explore advanced Explainable Artificial Intelligence techniques such as Score-CAM, XGrad-CAM, and attention-guided visualization methods. Incorporating segmentation-based frameworks such as U-Net or Mask R-CNN can also improve tumour boundary localization accuracy. In addition, lightweight deep learning architectures and model compression techniques may reduce computational complexity and support real-time deployment in clinical systems.

Future work may further investigate multimodal medical image fusion, federated learning for privacy-preserving healthcare analysis, and cloud-assisted AI frameworks for scalable and collaborative brain tumour diagnosis. These advancements can improve diagnostic reliability, interpretability, and practical clinical applicability of explainable deep learning systems for brain tumour detection.

7. Conclusion

This paper presented an explainable deep learning framework for brain tumour detection and classification using VGG16 and Grad-CAM on MRI brain images. The proposed framework effectively combines deep feature extraction capability with explainable visualization to improve both classification accuracy and interpretability. The VGG16 architecture successfully learned discriminative tumour features such as texture, intensity variations, and abnormal tissue patterns, while Grad-CAM generated visual heatmaps highlighting the important tumour regions responsible for classification decisions. Experimental analysis demonstrated that the proposed model achieved strong classification performance with an overall accuracy of 84.56%, along with reliable precision, recall, and F1-score values across multiple tumour categories. The confusion matrix and Grad-CAM visualizations further confirmed that the framework accurately identified medically relevant tumour regions with minimal inter-class misclassification. The explainability provided by Grad-CAM improves transparency and enhances clinician trust in automated diagnostic systems. The proposed framework offers several advantages, including automatic feature learning, improved interpretability, reduced dependency on manual diagnosis, and better support for clinical decision-making. Although the framework achieved promising results, performance may be further improved using larger MRI datasets, hybrid transformer-based architectures, and advanced attention mechanisms. Future work may also focus on multimodal MRI analysis, real-time clinical deployment, and integration with federated or cloud-based healthcare systems for more robust and scalable brain tumour diagnosis.

References

1. Nhlapho, Wandile, Marcellin Atemkeng, Yusuf Brima, and Jean-Claude Ndogmo. "Bridging the gap: exploring interpretability in deep learning models for brain tumor detection and diagnosis from MRI images." *Information* 15, no. 4 (2024): 182.
2. Li, Zhengkun, and Omar Dib. "Empowering brain tumor diagnosis through explainable deep learning." *Machine Learning and Knowledge Extraction* 6, no. 4 (2024): 2248-2281.
3. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
4. Tjoa, Erico, and Cuntai Guan. "A survey on explainable artificial intelligence (xai): Toward medical xai." *IEEE transactions on neural networks and learning systems* 32, no. 11 (2020): 4793-4813.

5. Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: visual explanations from deep networks via gradient-based localization." *International journal of computer vision* 128, no. 2 (2020): 336-359.
6. Ge, Chenjie, Irene Yu-Hua Gu, Asgeir Store Jakola, and Jie Yang. "Enlarged training dataset by pairwise GANs for molecular-based brain tumor classification." *IEEE Access* 8 (2020): 22560-22570.
7. Polat, Özlem, Zümray Dokur, and Tamer Ölmez. "Brain tumor classification by using a novel convolutional neural network structure." *International Journal of Imaging Systems and Technology* 32, no. 5 (2022): 1646-1660.
8. Kataria, Pooja, Ayush Dogra, Mili Gupta, Tripti Sharma, and Bhawna Goyal. "Trends in DNN model based classification and segmentation of brain tumor detection." *The Open Neuroimaging Journal* 16, no. 1 (2023).
9. Shajin, Francis H., Salini P, Paulthurai Rajesh, and Venu Kadur Nagoji Rao. "Efficient framework for brain tumour classification using hierarchical deep learning neural network classifier." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 11, no. 3 (2023): 750-757.
10. Chukwujindu, Ezekiel, Hafsa Faiz, AI-Douri Sara, Khunsa Faiz, and Alexandra De Sequeira. "Role of artificial intelligence in brain tumour imaging." *European Journal of Radiology* 176 (2024): 111509.
11. Guluwadi, Suresh. "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50." *BMC Medical Imaging* 24, no. 1 (2024): 1-19.
12. Haque, Rezuana, Md Mehedi Hassan, Anupam Kumar Bairagi, and Sheikh Mohammed Shariful Islam. "NeuroNet19: an explainable deep neural network model for the classification of brain tumors using magnetic resonance imaging data." *Scientific Reports* 14, no. 1 (2024): 1524.
13. Basha, Niha Kamal, Christo Ananth, K. Muthukumaran, Gadug Sudhamsu, Vikas Mittal, and Fikreselam Gared. "Mask region-based convolutional neural network and VGG-16 inspired brain tumor segmentation." *Scientific Reports* 14, no. 1 (2024): 17615.
14. Zeineldin, Ramy A., Mohamed E. Karar, Ziad Elshaer, Jan Coburger, Christian R. Wirtz, Oliver Burgert, and Franziska Mathis-Ullrich. "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI." *Scientific Reports* 14, no. 1 (2024): 3713.
15. Amarneni, Sateesh, and R. S. Valarmathi. "Diagnosing the MRI brain tumour images through RNN-LSTM." *e-Prime-Advances in Electrical Engineering, Electronics and Energy* 9 (2024): 100723.
16. Chel, Yoon Han, and Lin Lih Poh. "Brain tumor classification in MRI: Insights from LIME and Grad-CAM explainable AI techniques." *IEEE Access* (2025).
17. Mir, Aaqib Rashid, and Bachcha Lal Pal. "Interpretable brain tumor detection using VGG16 and grad-CAM visualization." *International Journal of Scientific Research* 14 (2025): 927-932.
18. Rao, Ch Dhanunjaya, and A. Kumaravel. "Hybrid DNN-VGG16 Framework for Brain Tumor Segmentation and Classification Using MRI Imaging: an Optimized Deep Learning Approach." In *International Conference on Intelligent Computing and Communication*, pp. 599-614. Cham: Springer Nature Switzerland, 2025.
19. Salomi, M., and Parth Nagrecha. "Comparative Analysis of Brain Tumor Classification Using DNN." In *2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, pp. 1-6. IEEE, 2025.
20. https://www.kaggle.com/datasets/khajaahmed1/brain-tumor-mri-dataset/data?utm_source