Early Brain Tumor Detection using XAI

E Mohan[#1], D. Bala Venkata Sai[#2], A Venkata Charan[#3], Mr. K. Saravanan[#4]
#[1,2,3] UG Student, Department of Computer Science and Engineering, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Trichy – 621112, Tamil Nadu, India
#[4] Assistant Professor, Department of Computer Science and Engineering, Dhanalakshmi Srinivasan University, Trichy – 621112, Tamil Nadu, India

***Abstract*— Brain tumors are among the most life-threatening neurological disorders, and their early detection plays a crucial role in improving patient survival rates and treatment planning. Recent advancements in Artificial Intelligence (AI), particularly deep learning techniques, have shown significant potential in automated medical image analysis. Convolutional Neural Networks (CNNs) are widely adopted for analyzing Magnetic Resonance Imaging (MRI) scans due to their capability to learn hierarchical and discriminative features. However, conventional deep learning models often function as black-box systems, limiting their interpretability and clinical trust. To address this limitation, this paper proposes an AI-based framework for early brain tumor detection using MRI images integrated with Explainable Artificial Intelligence (XAI) techniques. The proposed system utilizes a fine-tuned VGG-16 CNN architecture for accurate tumor classification across four categories and incorporates Gradient-weighted Class Activation Mapping (Grad-CAM) to generate visual explanations highlighting tumor-affected regions. Experimental results on the publicly available Kaggle Brain Tumor MRI Dataset demonstrate a classification accuracy of 96.5%, with Grad-CAM visualizations confirming clinically meaningful tumor localization, thereby supporting reliable and transparent clinical decision-making.**

Keywords*— Artificial Intelligence, Brain Tumor Detection, MRI Images, Deep Learning, CNN, Grad-CAM, Explainable AI, XAI, VGG-16, Healthcare, Medical Imaging.

I. INTRODUCTION

Brain tumors represent one of the most dangerous and rapidly progressing neurological conditions affecting patients worldwide. According to global cancer statistics, approximately 308,000 new cases of primary brain and central nervous system tumors are diagnosed annually, with a five-year survival rate below 36% for malignant gliomas when detected in advanced stages [1]. Early and accurate diagnosis is therefore the single most critical determinant of patient outcomes.

Magnetic Resonance Imaging (MRI) is the clinical gold standard for brain tumor detection owing to its superior soft-tissue contrast and ability to capture detailed structural information without ionizing radiation. However, the increasing volume of MRI data has created strong demand for automated, intelligent systems capable of analyzing scans efficiently and consistently, reducing the burden of manual interpretation and inter-observer variability.

Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in medical image analysis by learning hierarchical and spatial features directly from raw image data without manual feature engineering [2]. CNN-based models have been applied to brain tumor detection, segmentation, and classification with high diagnostic accuracy. Despite their effectiveness, most deep learning models operate as black-box systems, providing predictions without explaining the underlying reasoning — a critical limitation in healthcare where clinicians require transparency to trust and validate AI-assisted diagnoses.

Explainable Artificial Intelligence (XAI) addresses this challenge by enabling deep learning models to generate human-interpretable explanations for predictions. Gradient-weighted Class Activation Mapping (Grad-CAM) [3] is a widely adopted XAI technique that produces class-discriminative heatmaps visually highlighting image regions that most influence model decisions. By integrating Grad-CAM into the classification pipeline, the proposed system bridges the gap between predictive performance and clinical trust.

This paper presents an AI-based brain tumor detection system integrating a fine-tuned VGG-16 CNN with Grad-CAM XAI. The system classifies MRI scans into four categories — Glioma, Meningioma, Pituitary Tumor, and No Tumor — and generates heatmap overlays to localize tumor-affected regions. The remainder of this paper is organized as follows: Section II presents related work; Section III describes the proposed system; Section IV elaborates the methodology; Section V details the architecture; Sections VI and VII present results and comparative analysis; Section VIII concludes; and Section IX discusses future work.

II. RELATED WORK

Deep learning-based approaches to medical image classification have advanced significantly over the past decade. Litjens et al. [2] conducted a comprehensive survey demonstrating that CNN-based methods consistently outperform traditional machine learning approaches across detection, segmentation, and classification tasks in medical imaging.

In clinical diagnostics, Rajpurkar et al. [4] proposed CheXNet, a deep CNN achieving radiologist-level pneumonia detection from chest X-rays, while Esteva et al. [5] demonstrated dermatologist-level skin cancer classification using deep neural networks. These works established the clinical viability of deep learning but relied on black-box models without interpretability mechanisms.

For brain tumor classification specifically, Afshar et al. proposed capsule networks achieving 90.89% accuracy on a three-class MRI dataset, though without segmentation or explainability support. Cheng et al. employed VGG-16 features with SVM classifiers reporting 95.6% accuracy, but lacked visual explanation modules critical for clinical trust.

Selvaraju et al. [3] introduced Grad-CAM as a model-agnostic visualization technique producing class-discriminative heatmaps compatible with standard CNN architectures. Grad-CAM has since been widely adopted in medical imaging for post-hoc explainability. Complementary techniques including SHAP [8] and LIME [7] offer alternative explanation paradigms but introduce significant computational overhead compared to Grad-CAM.

Existing systems predominantly address either classification accuracy or visual explainability in isolation, without integrating both in a unified deployable framework. The proposed system addresses this gap by combining high-accuracy VGG-16 classification with efficient Grad-CAM explainability in a cohesive pipeline designed for practical clinical deployment.

TABLE I: Comparison of Existing Methods

Ref.	Method	Dataset	Acc.	Limitation
Pereira	CNN Seg.	BraTS'15	0.78 DSC	No XAI
Afshar	CapsNet	Figshare	90.9 %	No Expl.
Cheng	VGG+SVM	Public	95.6 %	No XAI
Saeedi	Att-UNet	BraTS'21	0.90 DSC	No Deploy
Proposed	VGG16+GradCAM	Kaggle	96.5 %	None

III. PROPOSED SYSTEM

The proposed system is an XAI-based framework for early brain tumor detection using MRI images. The system accepts brain MRI scans and classifies them into four categories: Glioma, Meningioma, Pituitary Tumor, and No Tumor. Simultaneously, Grad-CAM heatmaps are generated to visually localize tumor-affected regions within the MRI scans, providing transparent and interpretable AI-assisted diagnoses suitable for clinical deployment.

The core classification backbone is a VGG-16 network pre-trained on ImageNet and fine-tuned on the Kaggle Brain Tumor MRI Dataset. The standard VGG-16 fully connected head is replaced with a custom classification head comprising Global Average Pooling, a Dense layer (512 units, ReLU), Dropout (rate=0.5), and a Softmax layer with four output neurons. The model is trained end-to-end using the Adam optimizer with categorical cross-entropy loss.

Grad-CAM is applied to the trained model by computing the gradient of the predicted class score with respect to feature maps at the final convolutional block (block5_conv3). These gradients are globally average-pooled to derive neuron importance weights, which are combined with forward-pass activation maps and upsampled to input resolution. The resulting heatmap is overlaid on the original MRI scan in a jet colormap, enabling radiologists to verify model focus areas correspond to actual tumor regions.

The combined output — predicted class label, confidence score, and Grad-CAM heatmap overlay — provides clinicians with both quantitative and visual diagnostic insights. This design ensures the system is not merely a black-box predictor but a transparent decision-support tool that enhances clinical confidence and supports radiologist validation of AI predictions.

IV. METHODOLOGY

A. Data Collection

The Kaggle Brain Tumor MRI Dataset is used as the primary training and evaluation resource, containing 7,023 labeled MRI images across four classes: Glioma (1,621), Meningioma (1,645), No Tumor (2,000), and Pituitary Tumor (1,757). All images are T1-weighted contrast-enhanced MRI scans in JPEG format. The dataset is partitioned into 80% training (5,712 images) and 20% testing (1,311 images) subsets.

B. Image Preprocessing

All MRI images are resized to 224×224 pixels to match the VGG-16 input requirement. Pixel values are normalized to [0, 1] by dividing by 255. Data augmentation during training includes random horizontal flip (p=0.5), rotation (±15°), zoom (±10%), and width/height shift to improve generalization and reduce overfitting.

C. Feature Extraction and Classification

A VGG-16 backbone initialized with ImageNet weights is fine-tuned end-to-end. The custom head comprises: Global Average Pooling → Dense (512, ReLU) → Dropout (0.5) → Softmax (4 classes). Training uses Adam optimizer (lr=0.0001), categorical cross-entropy loss, batch size of 32, for 30 epochs with early stopping (patience=5) based on validation loss.

D. Explainability via Grad-CAM

Post-inference, Grad-CAM computes gradients of the predicted class score with respect to block5_conv3 feature maps of VGG-16. Gradients are globally average-pooled into neuron importance weights, linearly combined with activation maps, ReLU-activated, and upsampled to 224×224. The heatmap is rendered with a jet colormap and superimposed on the original MRI at 50% transparency.

E. Performance Evaluation

The system is evaluated on Accuracy, Precision, Recall, F1-Score per class, and Inference Time per image. Qualitative assessment of Grad-CAM heatmaps is conducted by visual inspection to verify alignment of highlighted regions with anatomically relevant tumor locations.

V. SYSTEM ARCHITECTURE

The end-to-end architecture of the proposed XAI-based brain tumor detection system comprises six sequential processing stages as illustrated in Fig. 1. The pipeline flows from raw MRI image acquisition through preprocessing, VGG-16 feature extraction and classification, tumor class prediction, Grad-CAM heatmap generation, and final output delivery to the clinician.

Stage 1 (Input): Brain MRI images from the Kaggle dataset are ingested in JPEG format. Stage 2 (Preprocessing): Images are resized to 224×224 px, normalized, and augmented. Stage 3 (CNN): The fine-tuned VGG-16 performs hierarchical feature extraction across 13 convolutional layers, followed by the custom classification head producing class probabilities via Softmax. Stage 4 (Prediction): The class with maximum probability and its confidence score are returned. Stage 5 (Grad-CAM): Gradients at block5_conv3 generate the class-specific heatmap. Stage 6 (Output): The predicted class, confidence score, and heatmap overlay are displayed to the radiologist for clinical review.

Fig. 1: XAI-Based Brain Tumor Detection Pipeline

INPUT: MRI Image Dataset (Kaggle – 7,023 images, 4 classes)
PRE-PROCESSING: Resize 224×224 | Normalize [0,1] Augmentation: flip, rotate, zoom
VGG-16 CNN: 13 Conv Layers + Custom Head GAP → Dense(512) → Dropout → Softmax(4)
PREDICTION: Class: Glioma / Meningioma / Pituitary / No Tumor + Confidence (%)
GRAD-CAM: Gradients @ block5_conv3 Class-specific Heatmap Generation
OUTPUT: Heatmap Overlay + Class + Confidence → Radiologist Review

Fig. 1: System Architecture — End-to-End XAI Pipeline

VI. RESULTS AND DISCUSSION

The proposed system was evaluated on the Kaggle Brain Tumor MRI Dataset test split (1,311 images). The VGG-16 model trained for 30 epochs achieved a classification accuracy of 96.5%, with strong per-class metrics across all four tumor categories as presented in Table II.

TABLE II: Per-Class Classification Performance

Tumor Class	Prec.	Rec.	F1	Sup.
Glioma	97.2	96.8	97.0	321
Meningioma	94.1	93.7	93.9	306
No Tumor	98.3	97.9	98.1	405
Pituitary	96.8	97.2	97.0	279
Overall	96.5	96.4	96.5	1311

The CNN model effectively captures discriminative spatial and intensity-based features from MRI scans, with minimal inter-class misclassification. The Meningioma class shows the lowest F1-score (93.9%) due to its morphological similarity with normal brain tissue, which is consistent with challenges reported in existing literature. Glioma and Pituitary Tumor achieve the highest performance owing to their visually distinctive appearance in T1-weighted contrast-enhanced MRI.

Grad-CAM visualizations confirm that model attention is consistently directed toward tumor-bearing regions in correctly classified MRI slices. In Glioma cases, heatmaps highlight ring-enhancing lesion boundaries; in Meningioma cases, peripheral dural attachment areas are activated; and in Pituitary cases, the sellar region is correctly highlighted. These findings validate that the model learns clinically meaningful features rather than spurious background correlations.

Grad-CAM integration adds only 0.03 seconds per image to the base inference time of 0.9 seconds, confirming practical suitability for real-time clinical deployment. The availability of visual explanations enables healthcare professionals to effectively validate AI-assisted diagnoses, enhancing trust and adoption.

Fig. 2: Grad-CAM Heatmaps — Glioma, Meningioma, Pituitary, No Tumor

VII. PERFORMANCE EVALUATION AND COMPARATIVE ANALYSIS

A. Experimental Setup

All experiments were conducted using Python 3.10, TensorFlow 2.12, and Keras on an NVIDIA Tesla T4 GPU (Google Colab). The VGG-16 backbone was initialized with ImageNet weights and fine-tuned end-to-end for 30 epochs. Adam optimizer was used with initial learning rate 0.0001, decayed by 0.1 after every 10 epochs. Batch size was 32. Early stopping (patience=5) was applied on validation loss to prevent overfitting.

B. Evaluation Metrics

Five metrics were used: Accuracy (overall correctness), Precision (positive predictive value), Recall (sensitivity), F1-Score (harmonic mean of precision and recall), and Inference Time (computational efficiency). These provide a comprehensive view of both classification performance and real-time clinical suitability.

C. Comparative Analysis

The proposed system was compared against three baselines — Standard CNN (no XAI), CNN+LIME, and CNN+SHAP — all trained and tested on identical dataset splits. As shown in Table III, the proposed CNN+Grad-CAM achieves the highest accuracy (96.5%) with full heatmap explainability and the fastest inference time among XAI-augmented methods.

TABLE III: Comparative Analysis

System	Acc.	F1	Time	XAI
Std. CNN	94.2%	94.0%	0.87s	None
CNN+LIME	93.8%	93.5%	2.10s	Partial
CNN+SHAP	94.0%	93.7%	1.90s	Partial
Proposed	96.5%	96.5%	0.93s	Full Heatmap

D. Discussion on Explainability Impact

The Grad-CAM visualizations validate model predictions by highlighting tumor-relevant regions, enabling cross-verification by radiologists. The explainability component reduces the risk of misdiagnosis from spurious model attention and significantly increases clinical confidence in AI-assisted diagnoses. The proposed system achieves the best trade-off between accuracy, explainability quality, and computational efficiency among all compared methods.

VIII. CONCLUSION

This paper presented an AI-based system for early brain tumor detection from MRI images, integrating a fine-tuned VGG-16 CNN with Grad-CAM explainability in a unified clinical framework. The system achieves a classification accuracy of 96.5% across four tumor classes on the Kaggle Brain Tumor MRI Dataset, with Grad-CAM heatmaps consistently localizing tumor-affected anatomical regions in a clinically meaningful manner.

The proposed framework addresses two critical limitations of existing approaches: high classification accuracy without interpretability, and XAI methods that sacrifice performance for explainability. By achieving state-of-the-art accuracy while providing full visual explanations at near-real-time inference speed (0.93s/image), the system demonstrates practical suitability for clinical decision support. The open modular design supports extension to additional tumor types, imaging modalities, and deployment environments.

IX. FUTURE WORK

Future enhancements include: (1) integration of LIME and SHAP alongside Grad-CAM for complementary multi-perspective explainability; (2) extension to 3D volumetric MRI analysis using 3D-CNN or Vision Transformer architectures for improved detection of small, multi-slice tumors; (3) deployment via a secure web interface integrated with hospital PACS systems for direct DICOM ingestion and radiologist reporting; (4) integration of Monte Carlo Dropout for uncertainty quantification, enabling the system to report prediction confidence intervals; and (5) multi-institutional prospective clinical validation across diverse MRI scanner protocols to confirm generalizability and real-world clinical impact.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Trichy, Tamil Nadu, India, for providing the computational resources and academic support required to conduct this research. The authors also acknowledge the open-access Kaggle Brain Tumor MRI Dataset contributors whose data made this work possible.

REFERENCES

[1] Global Cancer Observatory (GLOBOCAN), Brain and Central Nervous System Tumors, International Agency for Research on Cancer, WHO, 2023.
[2] G. Litjens et al., A survey on deep learning in medical image analysis, IEEE Trans. Med. Imaging, vol. 36, no. 1, pp. 1-18, Jan. 2017.
[3] R. R. Selvaraju et al., Grad-CAM: Visual explanations from deep networks via gradient-based localization, in Proc. IEEE ICCV, 2017, pp. 618-626.
[4] P. Rajpurkar et al., CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning, arXiv:1711.05225, 2017.
[5] A. Esteva et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature, vol. 542, no. 7639, pp. 115-118, 2017.
[6] D. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Muller, Explainable AI: Interpreting, explaining and visualizing deep learning, IEEE Signal Process. Mag., vol. 38, no. 3, pp. 39-48, 2021.
[7] M. T. Ribeiro, S. Singh, and C. Guestrin, Why should I trust you? Explaining the predictions of any classifier, in Proc. ACM SIGKDD, 2016, pp. 1135-1144.
[8] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in Advances in NeurIPS, 2017, pp. 4765-4774.
[9] A. Holzinger et al., What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923, 2017.
[10] Z. Zhou, X. Li, and R. He, Explainable deep learning for medical diagnosis, IEEE Access, vol. 8, pp. 190930-190940, 2020.
[11] J. Kim et al., Interpretable deep learning in medical imaging, IEEE J. Biomed. Health Inform., vol. 22, no. 4, pp. 1302-1311, 2018.
[12] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in Proc. IEEE CVPR, 2016, pp. 770-778.
[13] S. A. Pereira et al., Brain tumor segmentation using convolutional neural networks in MRI images, IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1240-1251, 2016.
[14] P. Afshar, A. Mohammadi, and K. N. Plataniotis, Brain tumor type classification via capsule networks, in Proc. IEEE ICIP, 2019, pp. 3129-3133.
[15] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.