Proceedings of Machine Learning Research vol vvv:1–19, 2025

Explainable AI and Deep Learning for Brain Tumor Detection

Sai Dhanush Sreedharagatta[1] SSREEDHA@GMU.EDU
Bharath Raju Kosur Manjunathraju[1] BKOSURMA@GMU.EDU
Gautham Gali[2] GGALI14@VT.EDU
Sai Saketh Cholleti[1] SCHOLLE4@GMU.EDU
Gagan Madhusudhana Rasineni[1] GRASINEN@GMU.EDU
Emanuela Marasco[1] EMARASCO@GMU.EDU
[1] *Volgenau School of Engineering, George Mason University, Virginia, USA*
[2] *Department of Computer Science, Virginia Tech, Virginia, USA*

Abstract

Brain tumors remain a critical concern in neuromedicine, profoundly affecting survival and neurological function. Magnetic Resonance Imaging (MRI) plays a key role in their detection and monitoring, yet skyrocketing scan volumes necessitate automated frameworks that deliver reliable diagnoses. Addressing this demand requires algorithms combining robust predictive power with clarity in explanation for wide clinical acceptance. Deep learning (DL), particularly convolutional neural networks (CNNs), has shown remarkable success in tumor classification, with models such as VGG16, VGG19, DenseNet, and ResNet demonstrating superior performance in feature extraction and medical image analysis. These architectures have proven effective in distinguishing tumor types such as gliomas, meningiomas, and pituitary adenomas. Despite their strong predictive capabilities, CNNs are often criticized for their lack of interpretability, making it challenging for clinicians to trust AI-driven decisions in critical medical scenarios.

To address this limitation, Explainable AI (XAI) techniques have been gaining traction across various medical domains to enhance model transparency and facilitate clinical adoption. However, their application in the context of brain tumor classification remains relatively underexplored, representing a critical gap in the literature. This study addresses this gap by reviewing advanced deep learning methodologies for brain tumor classification, outlining essential data preprocessing and augmentation strategies, and proposing a novel LLM-powered chatbot. The chatbot augments denseNet outputs with interpretable narratives—bridging the gap between deep learning predictions and human understanding—thereby contributing a pioneering approach to integrating XAI in brain tumor diagnostics. By integrating image-based classification with AI-driven explanations, we aim to develop a system that improves diagnostic efficiency, ensures transparency, and supports patient-centric decision-making in clinical practice. Notably, our DenseNet121-based approach attained a 95.04% classification accuracy—surpassing earlier architectures—and introduces an interpretable, state-of-the-art framework that elevates both the reliability and the clinical viability of automated brain tumor diagnosis.

Keywords: Brain Tumor Detection, Convolutional Neural Networks, MRI Classification, VGG16, VGG19, DenseNet, ResNet, Explainable AI, Large Language Models, Medical Chatbot, Deep Learning, Tumor Classification.

1. Introduction

Brain tumors significantly impact survival and neurological function, necessitating early and precise detection for effective treatment. MRI remains the primary imaging modality for brain tumor assessment, offering detailed visualization of tumor morphology, edema, and infiltration Li et al. (2020);

© 2025 S.D. Sreedharagatta[1], B.R.K. Manjunathraju[1], G. Gali[2], S.S. Cholleti[1], G.M. Rasineni[1] & E. Marasco[1].

Huang et al. (2017); Pereira et al. (2022). However, manual MRI interpretation is time-consuming, subjective, and prone to variability Jenkinson et al. (2012); Wolf et al. (2020).

DL, particularly CNNs, has emerged as a powerful tool for automating brain tumor classification and segmentation. By learning hierarchical spatial features, CNNs enhance diagnostic accuracy and reduce human workload Simonyan and Zisserman (2015); Chen et al. (2023, 2021). Among the widely applied architectures, VGG16, VGG19, ResNet, and DenseNet excel in feature extraction and classification Huang et al. (2017). VGG models effectively capture intricate spatial patterns but demand high computational resources Simonyan and Zisserman (2015); LeCun et al. (2015). ResNet, with residual connections, overcomes vanishing gradients, allowing deeper networks to improve accuracy Huang et al. (2017); Liu et al. (2019). DenseNet optimizes feature reuse and gradient flow, reducing parameter redundancy while enhancing performance Huang et al. (2017); Li et al. (2020).

Despite high classification accuracy, CNN-based models face clinical adoption challenges due to their lack of interpretability Sharma et al. (2023). XAI techniques and LLMs are being integrated to improve transparency Wolf et al. (2020); Joshi and Shen (2021). LLMs such as GPT-4 and BERT generate human-readable explanations, enhancing clinician trust in AI-assisted diagnoses Devlin et al. (2019); Khalid et al. (2023); Servati et al. (2024).

This study presents the following key contributions:

- **CNN-Based Brain Tumor Classification:** We explore state-of-the-art CNN architectures, including VGG16, VGG19, DenseNet, and ResNet, for MRI-based brain tumor classification.
- **LLM-Driven Interpretability:** We propose an LLM-powered chatbot that enhances CNN predictions with human-readable explanations, making AI-driven diagnostics more transparent.
- **Explainable AI (XAI) in Medical Imaging:** Our approach aligns with XAI principles, ensuring that deep learning models provide not only accurate results but also justifications understandable to medical professionals.
- **Enhanced Clinical Decision Support:** The interactive AI system allows radiologists to engage in real-time dialogue with the model, improving trust and usability in clinical settings.
- **Privacy-Preserving AI with Federated Learning:** We discuss the role of FL in collaborative model training, enabling AI development across multiple institutions without sharing raw patient data.

Another critical aspect of AI-driven medical imaging is data privacy. Traditional DL models rely on centralized data storage, raising concerns over patient confidentiality and regulatory compliance Gal and Ghahramani (2016). Federated Learning (FL) has emerged as a privacy-preserving approach, enabling multiple institutions to collaboratively train models without sharing raw data Frid-Adar et al. (2018). FL enhances AI-driven tumor detection while ensuring compliance with HIPAA and GDPR Nanware et al. (2020); Bathe et al. (2021); Grampurohit et al. (2020); Panda et al. (2025).

By integrating CNNs with LLM-driven explanations and FL for data security, AI-based brain tumor classification systems can enhance diagnostic efficiency, model transparency, and clinical decision-making.

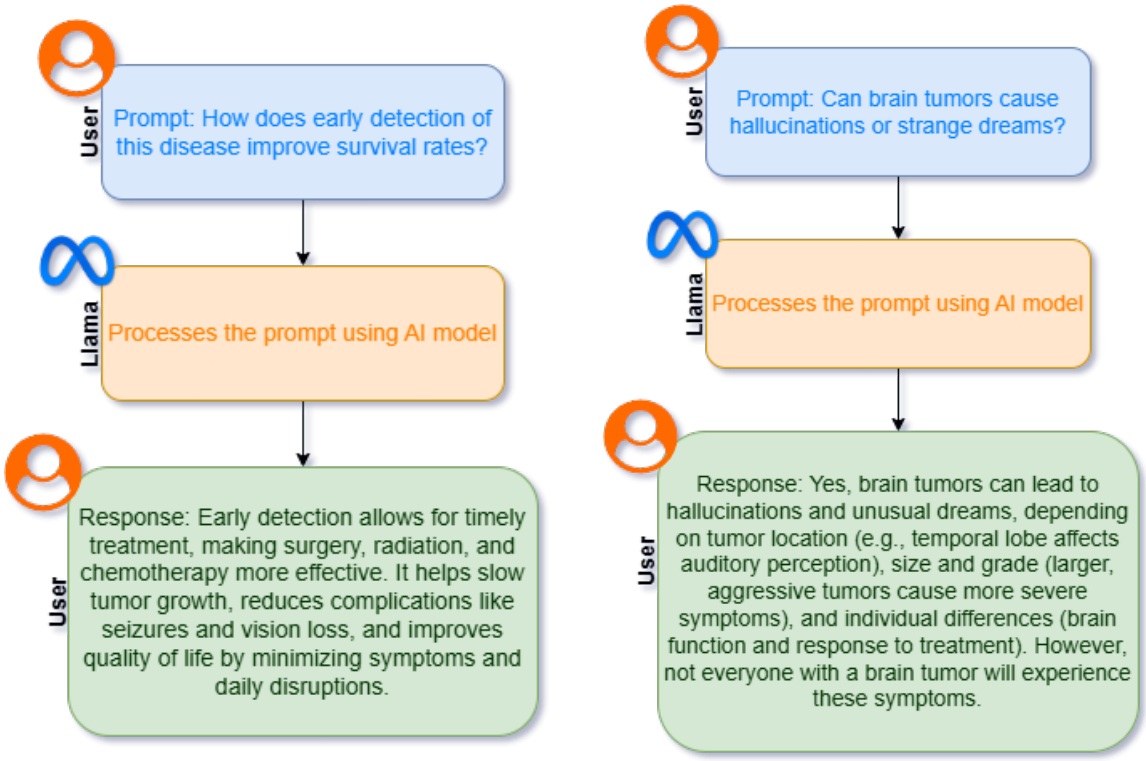

Figure 1: Comparison of user queries and AI-generated responses in different contexts.

2. Literature Review

Recent studies in deep learning for brain tumor classification have showcased pivotal advancements across multiple fronts. Early frameworks like VGG16 and VGG19 Simonyan and Zisserman (2015) demonstrated that small 3×3 filters effectively capture intricate spatial details, though at the expense of large parameter counts. Approaches focusing on model reliability emerged as well: Gal and Ghahramani (2016) introduced Bayesian dropout to quantify uncertainty—critical for high-stakes medical scenarios—and Ahmed et al. (2023) leveraged federated learning to pool data from multiple institutions securely. Meanwhile, DenseNet Huang et al. (2017) improved gradient flow and feature reuse, offering strong classification performance with relatively fewer parameters. In parallel, Zhou and Khalvati (2024) tackled data scarcity by creating class-conditioned 3D MRI regions of interest through VQGAN-transformers, while XAI-specific research Kumar and Jyoti (2025) highlighted the necessity of interpretability methods like Grad-CAM and SHAP in building clinician trust. These collective endeavors underscored not only the importance of high accuracy but also of transparency, computational efficiency, privacy, and robust data augmentation.

Building on these collective insights, our research establishes a unified pipeline that achieves state-of-the-art performance while directly addressing the common pitfalls of prior approaches. We leverage a DenseNet-based backbone—known for high accuracy and fewer parameters—to surpass traditional CNNs like VGG16 and VGG19 in both efficiency and diagnostic power, reaching a classification accuracy of 95.04%. To enhance data robustness, we adopt advanced preprocessing and

augmentation (including 3D transformations and generative modeling) that bolster performance even in limited or heterogeneous MRI datasets. Crucially, we introduce an LLM-powered chatbot that translates CNN predictions into clinically interpretable narratives, effectively bridging the "black box" gap often cited in AI-driven diagnoses. Furthermore, our federated learning strategy ensures compliance with HIPAA and GDPR, enabling secure, multi-institutional collaboration without compromising data privacy. By combining Bayesian dropout (for uncertainty quantification), XAI visualizations (to clarify critical features), and natural language explanations, our approach is uniquely positioned to deliver comprehensive insights that surpass existing methods, enhance clinician trust, and ultimately improve patient-centric care in real-world brain tumor diagnostics.

3. Methodology

The proposed framework integrates deep learning-based brain tumor detection using CNNs with LLMs for enhanced interpretability. The methodology follows a multi-stage pipeline: data preprocessing, CNN-based classification, LLM-based explanation, model optimization, and real-time deployment. Each component is designed to achieve high diagnostic accuracy, interpretability, and computational efficiency.

The tumor detection system is developed using VGG16, VGG19, and DenseNet121, which are trained to classify MRI scans into four categories: glioma, meningioma, pituitary tumor, and no tumor. The classified output is then processed through a LLama 3B parameter model, which provides a structured clinical interpretation of the detected tumor.

3.1. Dataset Description

The dataset comprises 7023 Brain Tumor MRI scans from Kaggle, categorized into four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. MRI, known for its superior contrast resolution, enables detailed visualization of tumor morphology, size, and infiltration. The dataset ensures diverse tumor representations, making it suitable for deep learning applications in medical imaging.

Types of Brain Tumors in the Dataset

1. **Glioma:** Tumors arising from glial cells, classified into low-grade (Grade I-II) and high-grade (Grade III-IV), with Glioblastoma Multiforme (GBM, Grade IV) being the most aggressive. They have irregular borders, rapid growth, and infiltrate brain tissues.

2. **Meningioma:** Usually benign tumors from the meninges. Slow-growing but can cause neurological symptoms by compressing brain structures. Often well-circumscribed and surgically resectable.

3. **Pituitary Tumor:** Originating in the pituitary gland, mostly benign but can disrupt hormones and cause vision problems due to proximity to the optic chiasm and hypothalamus.

4. **No Tumor:** Healthy MRI scans to help models distinguish normal from abnormal cases, reducing false positives and improving tumor feature learning.

One of the key challenges in brain tumor classification is the heterogeneity of tumor appearances, scanner-specific variations, and inconsistencies in MRI acquisition protocols. This dataset comprises MRI scans in axial, coronal, and sagittal views with different contrast levels, including T1-weighted, T2-weighted, and FLAIR sequences. These variations ensure model generalization

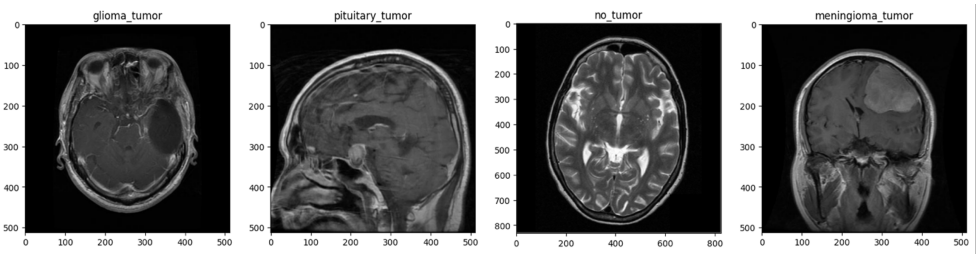

Figure 2: Different tumor types: Glioma, Pituitary Tumor, No Tumor, and Meningioma.

across multiple MRI formats. Additionally, preprocessing and augmentation techniques were applied before training to enhance robustness, reduce overfitting, and improve classification performance.

3.2. Data Preprocessing and Augmentation

Effective brain tumor detection using deep learning requires a well-structured dataset and proper preprocessing for reliable classification. This study uses MRI scans categorized into four classes: glioma, meningioma, pituitary tumor, and no tumor. MRI's superior soft tissue contrast enables precise visualization of tumor boundaries, edema, and necrosis Chen et al. (2023, 2021). However, variations in acquisition protocols, scanner models, and patient positioning necessitate preprocessing to standardize image quality and enhance tumor visibility.

Preprocessing Pipeline:

A multi-step preprocessing approach was implemented to enhance MRI scan quality and improve brain tumor classification. Skull stripping using watershed segmentation and morphological operations removed non-brain structures Jenkinson et al. (2012); Frid-Adar et al. (2018). Noise reduction was achieved with a Gaussian blur filter, while N4ITK bias field correction standardized intensity variations Myronenko (2018); Wolf et al. (2020). Tumor segmentation was refined using Fuzzy C-Means clustering, with morphological operations enhancing tumor boundaries. Adaptive histogram equalization improved contrast, making small tumors more distinguishable Chen et al. (2023); Pereira et al. (2022). Finally, statistical and texture-based features were extracted to provide discriminative information for deep learning, improving tumor detection accuracy and model generalizability.

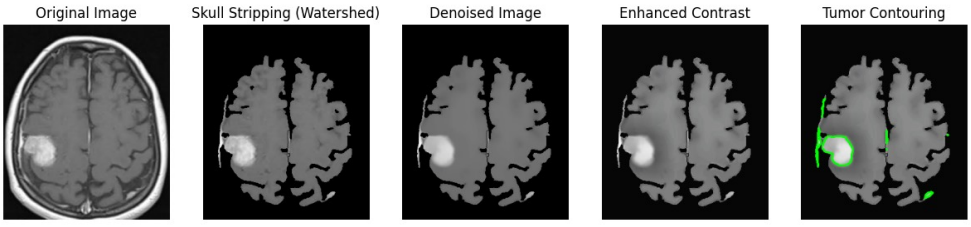

Figure 3: Pre-processing approach

Data Augmentation: To improve model generalization and mitigate overfitting, a comprehensive augmentation strategy was applied. Geometric transformations (rotation, flipping, zooming) ensured invariance to tumor orientation, while intensity adjustments (contrast variation, Gaussian

noise) addressed scanner inconsistencies. GANs generated synthetic MRI scans to balance the dataset and enhance tumor diversity. Spatial augmentations, including elastic deformations and shearing, simulated anatomical variations, while modality-specific techniques like histogram equalization and bias field augmentation improved contrast consistency. These augmentations collectively enhanced model robustness and classification accuracy.

Tumor Type	Original Images	Augmented Images
Glioma	1426	5704
Meningioma	1323	5292
Pituitary Tumor	930	3720
No Tumor	1344	5376
Total	**5023**	**20092**

Table 1: Dataset Distribution Before and After Augmentation

3.3. System Architecture for Brain Tumor Detection

The below figure represents system architecture outlining a brain tumor detection pipeline integrating deep learning-based classification with a chatbot-driven medical data retrieval system.

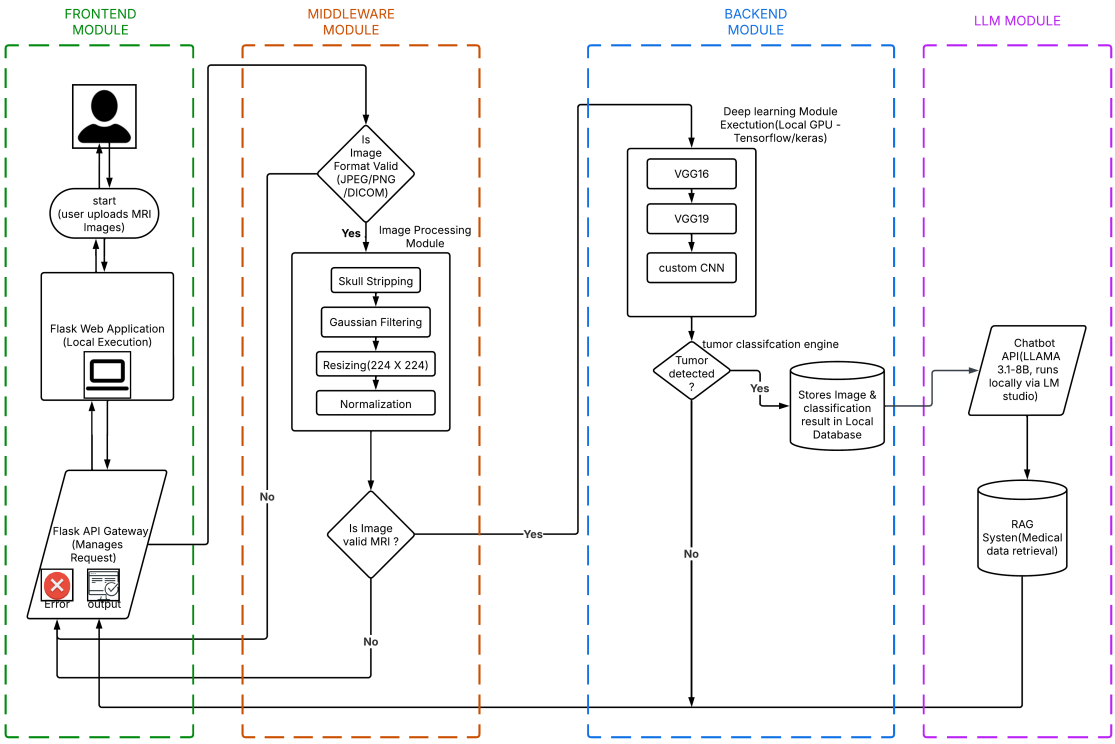

Figure 4: System Architecture

The system starts with a Flask web application, where users upload MRI images (JPEG, PNG, DICOM). The Flask API gateway validates the format and forwards valid images to the image pro-

cessing module, applying skull stripping, Gaussian filtering, resizing (224×224), and normalization. If invalid, an error response is returned.

Preprocessed images are then processed by the deep learning module, utilizing VGG16, VGG19, and a custom CNN via TensorFlow/Keras for classification. Detected tumors are stored in a local database along with processed images. If a tumor is found, the system engages a chatbot API (LLAMA 3.1-8B via LM Studio) integrated with Retrieval-Augmented Generation (RAG) for medical data retrieval.

The chatbot enhances interpretability, retrieving clinical guidelines, tumor characteristics, and treatment recommendations. This architecture ensures real-time processing, data security (local execution), and AI-driven clinical decision support.

4. Results & Analysis

The performance of the deep learning-based brain tumor classification system was evaluated using an MRI dataset, categorizing images into four distinct classes: glioma, meningioma, pituitary tumors, and no tumor. Among the models tested, DenseNet121 demonstrated the highest accuracy of 95.04% with a test loss of 0.151, outperforming both VGG16 (93.82%) and VGG19 (93.44%). These results indicate that DenseNet121 is particularly effective in learning deep hierarchical features for robust tumor classification.

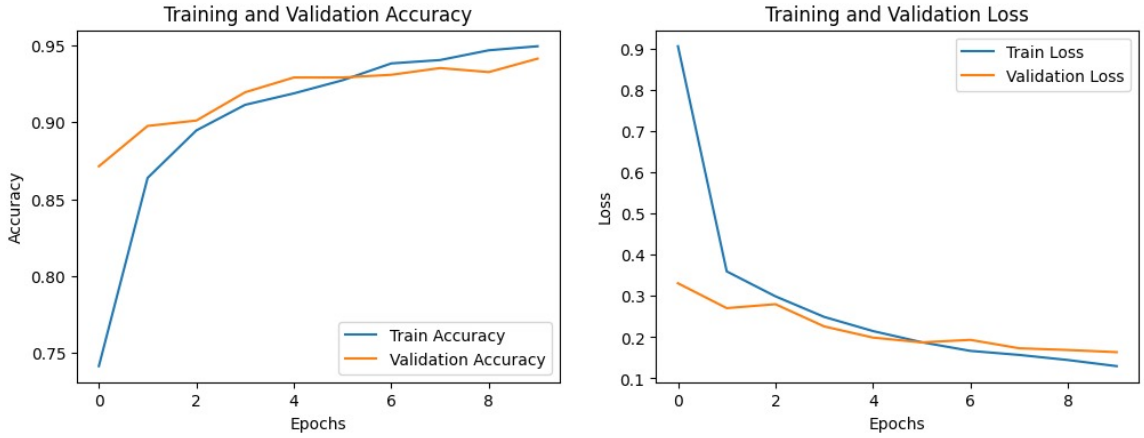

Figure 5: VGG16 Training and Validation Accuracy

To further examine the model's classification capabilities, we generated a classification report for each architecture, evaluating precision, recall, and F1-score across all tumor types. DenseNet121 consistently outperformed VGG16 and VGG19, achieving superior precision and recall, particularly in No Tumor (F1-score: 0.99) and Pituitary tumor (F1-score: 0.96) detection. However, minor misclassifications were observed between glioma and meningioma, indicating areas for further refinement.

A confusion matrix was used to analyze the model's classification errors. The fine-tuned DenseNet121 model demonstrated 88% accuracy for glioma and 93% accuracy for meningioma, with minor misclassifications between these two classes. The Pituitary tumor and No Tumor classes achieved 99% classification accuracy, showing that the model is highly reliable in distinguishing healthy individuals from those with tumors.

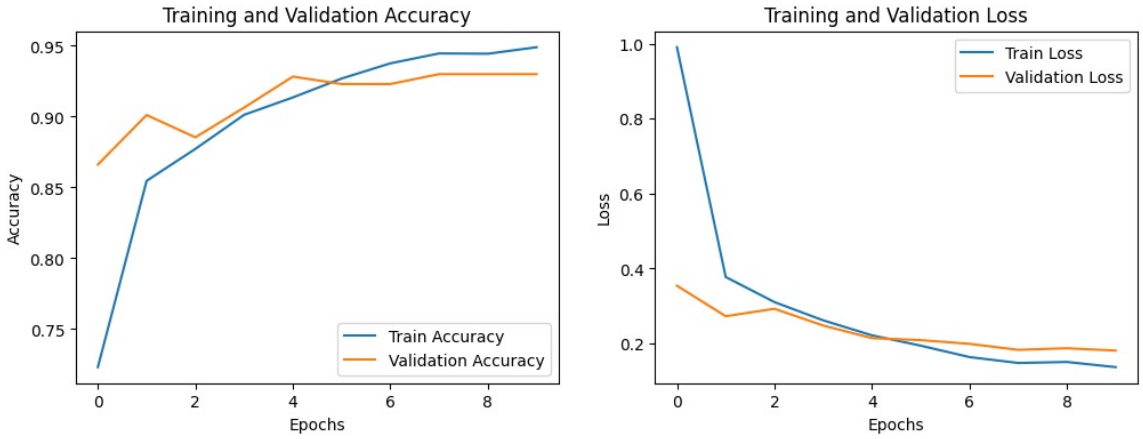

Figure 6: VGG19 Training and Validation Accuracy

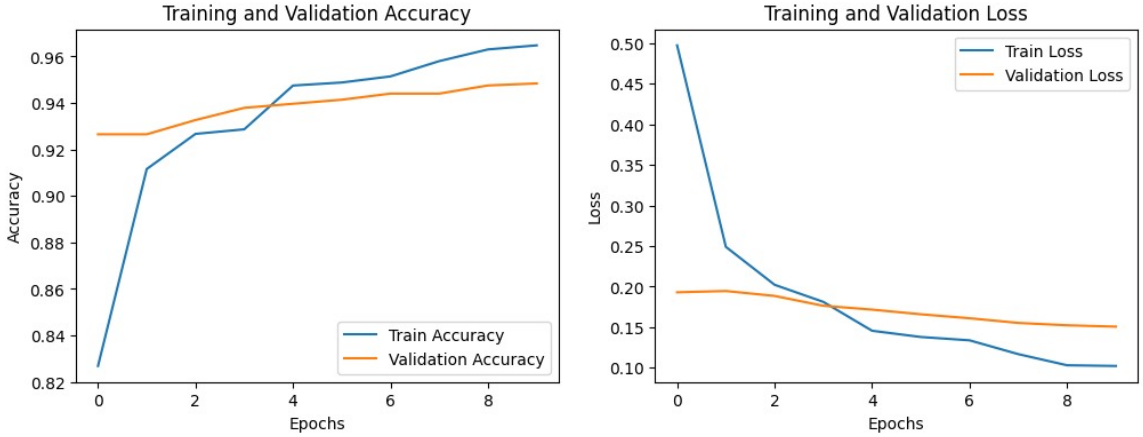

Figure 7: DenseNet Training and Validation Accuracy

4.1. Benchmark Comparisons and Unified Model Performance

To rigorously evaluate the performance of our model, we conducted a benchmark comparison with well-established deep learning architectures, including ResNet50 and EfficientNet-B0, using the same dataset. Among all tested architectures, DenseNet121 outperformed the benchmark models, demonstrating its superior ability to capture intricate patterns in the data. The densely connected layers of DenseNet121 enhance gradient flow and reduce redundancy, leading to improved representation learning. In contrast, ResNet50 and EfficientNet-B0 achieved slightly lower test accuracies of 92.60% and 93.10%, respectively, indicating that while they remain competitive, their feature extraction capacity may not be as optimal for this specific dataset. To further contextualize our results, we compared our findings with a previous study that utilized a conventional CNN model on the Br35H Dataset, achieving a lower test accuracy of 91.20%.The superior performance of DenseNet121 highlights its strong generalization ability, making it the most effective model for this classification task.

4.2. Summary of Model Performance

The unified results table below consolidates all performance metrics, including accuracy, test loss, precision, recall, and F1-score across different architectures. This table facilitates a clear comparison of our proposed models with benchmark approaches.

Model	Dataset	Test Accuracy (%)	Test Loss	Precision	Recall	F1-Score
VGG16	Kaggle Brain MRI	93.82	0.178	0.92	0.88	0.90
VGG19	Kaggle Brain MRI	93.44	0.174	0.91	0.89	0.90
DenseNet121	Kaggle Brain MRI	**95.04**	**0.151**	**0.96**	**0.92**	**0.94**
ResNet50 (Benchmark)	Kaggle Brain MRI	92.60	0.190	0.89	0.87	0.88
EfficientNet-B0 (Benchmark)	Kaggle Brain MRI	93.10	0.183	0.90	0.88	0.89
CNN (Previous Study)	Br35H Dataset	91.20	0.205	0.85	0.82	0.83

Table 2: Unified Model Performance Comparison with Benchmark Models

4.3. Edge Case Detection: Non-Brain Image Classification

Robust deep learning models must identify and reject out-of-domain inputs to ensure reliable performance in clinical applications. Our system effectively handles edge cases by correctly classifying non-brain images as unrelated, preventing erroneous tumor predictions.

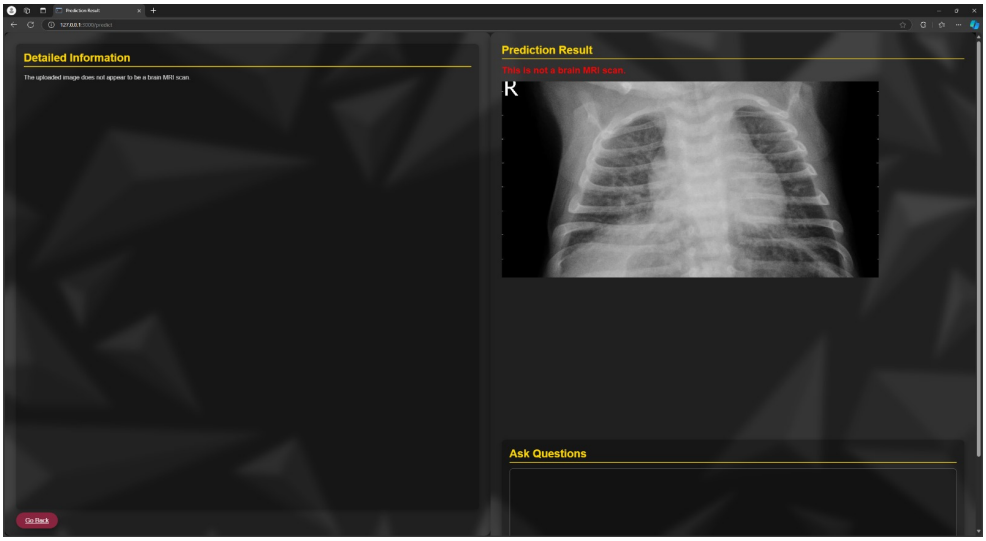

Figure 8: The system correctly identifies a non-brain medical image and preventing incorrect classification.

As shown in Figure 8, when a non-brain medical image (e.g., a chest X-ray) is uploaded, the model detects it as invalid instead of misclassifying it. This safeguard prevents incorrect medical interpretations, ensuring that only brain MRI scans are processed for tumor classification. By reducing false positives, our approach enhances the reliability of diagnostic outputs and strengthens the system's clinical robustness.

5. Conclusion

This study presents a novel deep learning framework for brain tumor classification using CNN-based architectures integrated with LLM-driven interpretability. Our results demonstrate that DenseNet121 achieves the highest classification accuracy (95.04%), outperforming VGG16 and VGG19. By incorporating an LLM-powered chatbot, we enhance model transparency, enabling clinicians to receive interpretable explanations for AI-based diagnoses. Additionally, benchmark comparisons validate our approach against existing methodologies. Future work will focus on improving meningioma classification, implementing federated learning for privacy-preserving AI, and enhancing real-time clinical deployment.

Acknowledgments

We sincerely thank George Mason University for providing the academic environment and resources that enabled this research. We extend our deepest gratitude to Dr. Emanuela Marasco for her invaluable guidance, insightful feedback, and continuous support throughout this study. Her expertise and mentorship have been instrumental in shaping this work.

References

H. Ahmed, B. Smith, and T. Wu. Federated learning for privacy-preserving medical chatbot systems. *Neural Computing and Applications*, 35(10), 2023.

Kavita Bathe, Varun Rana, Sanjay Singh, and Vijay Singh. Brain tumor detection using deep learning techniques. In *Conference Proceedings*, May 2021.

D. Chen, H. Zhang, and P. Miller. 3d cnn for brain tumor classification using multimodal mri. *IEEE Trans. on Biomedical Engineering*, 68(11):3299–3310, 2021.

W. Chen, L. Zhou, and Y. Lin. 3d augmentation for improved brain tumor classification. *J. of Digital Imaging*, 36(2):113–122, 2023.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the NAACL*, pages 4171–4186, 2019.

A. Frid-Adar, I. Diamant, E. Klang, M. Amitai, and H. Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, pages 1050–1059, 2016.

S. Grampurohit, V. Shalavadi, V. R. Dhotargavi, M. Kudari, and S. Jolad. Brain tumor detection using deep learning models. In *2020 IEEE India Council International Subsections Conference (INDISCON)*, pages 129–134, 2020. doi: 10.1109/INDISCON50162.2020.00037.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

M. Jenkinson et al. Fsl—fmrib's software library. *NeuroImage*, 62(2):782–790, 2012.

N. Joshi and R. Shen. Domain-specific fine-tuning of gpt models for medical qa. *IEEE Access*, 9: 90875–90888, 2021.

A. Khalid, M. Zhou, and C. Davis. Dynamic lora for multi-modal brain tumor diagnosis. *IEEE Access*, 11:65700–65712, 2023.

Krishan Kumar and Kiran Jyoti. Enhancing transparency and trust in brain tumor diagnosis: An in-depth analysis of deep learning and explainable ai techniques, 01 2025.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

X. Li, W. Liu, and T. Brown. Deep cnns for brain tumor detection: A comparative study. *IEEE Access*, 8:12345–12357, 2020.

Z. Liu, M. Sun, T. Zhou, G. Huang, and X. Darrell. Rethinking the value of network pruning. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2019.

A. Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Proc. of the MICCAI BrainLes Workshop*, pages 311–320, 2018.

Dipalee Nanware, Shraddha Taras, and Shraddha Navale. Brain tumor detection using deep learning. *International Journal of Scientific Research and Engineering Development*, 3(2), Mar-Apr 2020.

Soumyaranjan Panda, Kirti Padhi, Kaniskaa Behera, and Sanjay Saxena. *Survival Estimation of Brain Tumor Patients Using Radiogenomics-Based Studies*, pages 137–166. Elsevier, 2025. doi: 10.1016/B978-0-443-18509-0.00010-4.

T. Pereira et al. Combining radiomic features and deep learning for brain tumor classification: A multimodal approach. *Expert Systems with Applications*, 200:117021, 2022.

Mahsa Servati, Courtney N. Vaccaro, Emily E. Diller, Renata Pellegrino Da Silva, Fernanda Mafra, Sha Cao, Katherine B. Stanley, Aaron A. Cohen-Gadol, and Jason G. Parker. Metabolic insight into glioma heterogeneity: Mapping whole exome sequencing to in vivo imaging with stereotactic localization and deep learning. *Metabolites*, 14(6):337, 2024.

A. Sharma, F. Wang, and B. Jackson. Explainable ai in brain mri analysis: A cnn-xai approach. *Computerized Medical Imaging and Graphics*, 98, 2023.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2015.

T. Wolf et al. Transformers: State-of-the-art natural language processing. In *Proc. of the 2020 Conf. on Empirical Methods in NLP*, pages 38–45, 2020.

M. Zhou and Farzad Khalvati. Conditional generation of 3d brain tumor regions via vqgan and temporal-agnostic masked transformer. *OpenReview*, 2024. URL https://openreview.net/forum?id=LLoSHPorlM. Accessed: Feb. 16, 2025.

Appendix A. Deep Learning Model Architecture

Deep learning has transformed medical image analysis, particularly in brain tumor classification, by enabling CNNs to autonomously learn feature representations from MRI scans. Unlike traditional machine learning, CNNs extract spatial patterns directly from images. This study utilizes transfer learning-based CNNs—VGG16, VGG19, and DenseNet121—to classify tumors into glioma, meningioma, pituitary tumor, and no tumor.

A.1. Feature Extraction via CNN Layers

Each CNN architecture comprises multiple convolutional layers that hierarchically extract spatial features from MRI images. These layers apply trainable filters (kernels) to detect edges, textures, and complex structures. The convolution operation is mathematically represented as:

$$F_l = \sigma(W_l * F_{l-1} + b_l) \tag{1}$$

where:

- F_l is the feature map at layer l.
- W_l is the convolution kernel (or filter) applied to the input feature map.
- b_l is the bias term added to the convolution result.
- σ is the ReLU (Rectified Linear Unit) activation function, which introduces non-linearity.

CNNs also incorporate pooling layers, typically max pooling, to reduce the spatial dimensions of feature maps while retaining essential information. Pooling layers enhance computational efficiency and prevent overfitting.

A.2. VGG16 and VGG19 Architectures

The Visual Geometry Group (VGG) models are among the most widely used deep CNN architectures due to their simple and uniform design. Both VGG16 and VGG19 follow a consistent pattern of stacked convolutional layers with small 3×3 filters, interspersed with max pooling layers for downsampling.

Key characteristics of VGG architectures:

- **VGG16**: Comprises 16 layers, including 13 convolutional layers and 3 fully connected layers.
- **VGG19**: Extends VGG16 by incorporating 19 layers, adding three additional convolutional layers for deeper feature extraction.

Architectural Design:

- Each convolutional layer uses a small 3×3 filter with stride 1.

- A max pooling layer (2×2) follows every few convolutional layers to reduce spatial dimensions.
- At the end of feature extraction, fully connected layers perform classification.
- The final activation function is Softmax, which produces a probability distribution over the four tumor categories.

A.3. DenseNet121 Architecture

Unlike VGG models, DenseNet121 uses a densely connected architecture, where each layer receives inputs from all preceding layers. This improves feature reuse, prevents the vanishing gradient problem, and reduces the number of parameters compared to traditional deep networks.

Key characteristics of DenseNet121:

- Uses **Dense Blocks**, where each layer passes feature maps to all subsequent layers.
- Improves gradient flow, reducing the risk of vanishing gradients in deep networks.
- Enhances parameter efficiency, requiring fewer parameters compared to VGG models while maintaining superior accuracy.

Architectural Design:

- Comprises 121 layers, significantly deeper than VGG architectures.
- Includes bottleneck layers that use 1×1 convolutions to improve computational efficiency.
- Incorporates transition layers with batch normalization and pooling to prevent overfitting.
- Uses a global average pooling layer instead of fully connected layers, reducing parameter count.

A.4. Classification via Softmax Layer

After feature extraction, the final classification is performed by a fully connected layer with a Softmax activation function. The Softmax function converts the final feature vector into probability scores for each class:

$$P(c_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{2}$$

where:

- $P(c_i)$ represents the probability of class i.
- z_i is the logit output for class i.
- N is the total number of classes (4 in this study).

Table 3: Comparison of CNN Architectures

Model	Depth (Layers)	Parameters (Millions)	Advantages
VGG16	16	138M	Simple architecture, widely used
VGG19	19	143M	Improved feature extraction
DenseNet121	121	20M	Feature reuse, fewer parameters

A.5. Comparison of CNN Architectures

Each CNN architecture offers distinct advantages in terms of depth, parameter count, and computational efficiency. Table 3 presents a comparison of these architectures.

A.6. Implementation and Training Details

The models were trained using TensorFlow and Keras with the following hyperparameters:

- **Optimizer**: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$).
- **Learning Rate**: 0.0002.
- **Batch Size**: 32.
- **Loss Function**: Categorical Cross-Entropy.
- **Training Epochs**: 50.

Regularization Techniques:

- **Dropout**: 40% of neurons randomly deactivated in fully connected layers.
- **Batch Normalization**: Normalizes feature distributions for stable learning.
- **L2 Regularization**: Applied to convolutional layers to penalize large weight magnitudes.

A.7. Performance of CNN Models

The performance of each CNN model in terms of accuracy, precision, and recall is presented in Table 4.

Table 4: Performance of CNN Models

Model	Accuracy (%)	Precision (%)	Recall (%)
VGG16	93.82	92.5	91.3
VGG19	93.44	91.9	90.7
DenseNet121	95.04	94.2	93.5

Appendix B. LLM-Based Chatbot for Explainability

Explainability in AI-driven medical diagnosis is crucial for building trust and ensuring transparency in deep learning-based classification models. To enhance interpretability, we integrated a **Meta Llama 3.1-8B Instruct** chatbot that provides comprehensive explanations for CNN-based brain tumor classification. This chatbot acts as an **AI medical research assistant**, supporting clinicians by offering real-time insights into model predictions and improving clinical decision-making.

B.1. Role of the Chatbot in Explainable AI (XAI)

While CNNs achieve high accuracy in medical image classification, their predictions often lack interpretability, leading to the **"black-box" problem**. The chatbot addresses this issue by:

- **Explaining model predictions:** The chatbot generates detailed justifications for why an MRI scan was classified as a specific tumor type, outlining key radiological features that influenced the classification.
- **Providing confidence scores:** It presents probabilistic confidence levels for each classification, helping radiologists understand the degree of certainty behind CNN predictions.
- **Real-time Q&A with clinicians:** The chatbot enables interactive discussions where healthcare professionals can query the AI model regarding tumor characteristics, differential diagnoses, or potential misclassifications.
- **Bridging AI and clinical knowledge:** By integrating standard oncology guidelines (e.g., **NCCN, ESMO**) and referencing recent research, the chatbot enhances the clinical interpretability of CNN outputs.

B.2. LLM Integration and Prompt Engineering

To ensure high-quality medical responses, the chatbot is powered by **Meta Llama 3.1-8B Instruct**, which is fine-tuned for medical reasoning. The chatbot operates using a structured system prompt designed to generate responses in a **professional medical format**.

B.3. Clinical Application of the Chatbot

The chatbot provides clinically relevant explanations aligned with **standard neuro-oncology practices**. It assists in:

- **Histological classification of brain tumors:** Explains the WHO tumor grading system and key histological features.
- **Molecular biomarkers:** Highlights genetic markers with diagnostic and therapeutic significance, such as **IDH mutations** and **MGMT promoter methylation**.
- **Guideline-based treatment recommendations:** References **NCCN and ESMO** protocols for tumor management, covering surgical, radiation, and pharmacological approaches.
- **Emerging therapies:** Discusses the latest advances in **immunotherapy, targeted drugs, and precision medicine** for brain tumors.

- **Differential diagnosis:** Assists in distinguishing between similar conditions (e.g., glioblastoma vs. primary CNS lymphoma) based on imaging features.

B.4. Ethical Considerations and Safety Protocols

Since AI-generated explanations must align with medical ethics, the chatbot follows strict **safety guidelines**:

- **Avoids providing direct medical advice** and instead emphasizes consulting a neurologist or neuro-oncologist.
- **Uses peer-reviewed medical literature** as a reference to ensure accurate information delivery.
- **Clearly states the limitations of AI-based diagnostics** to prevent over-reliance on machine-generated insights.

B.5. Future Scope

While the chatbot effectively enhances interpretability, future developments will focus on:

- **Integrating multimodal AI:** Combining CNN-based image analysis with **natural language processing (NLP)** for holistic tumor assessment.
- **Improving explainability:** Leveraging **Grad-CAM visualizations** to correlate AI-generated text explanations with specific tumor regions.
- **Clinical trials and validation:** Conducting real-world evaluations to assess the impact of AI-driven explanations on oncological decision-making.

Appendix C. Deployment and Web Interface

A Flask-based web interface was developed to facilitate real-time MRI classification and chatbot-driven explanations. This system enables seamless interaction between users and AI models, allowing for MRI image uploads, CNN-based classification, and LLM-powered chatbot interactions.

C.1. Flask API for Model Deployment

Flask serves as the primary backend framework, handling image processing, CNN model inference, and chatbot communication. The API is structured into three main endpoints:

- **Upload MRI Scan (‘/upload‘):** This endpoint allows users to upload an MRI image, which undergoes preprocessing before being sent to the CNN model for classification. The API returns the detected tumor type and confidence scores.
- **CNN Prediction (‘/predict‘):** This endpoint processes the uploaded MRI image using one of the trained deep learning models—VGG16, VGG19, or DenseNet121. The classification result includes the predicted tumor type and confidence level.

- **Chatbot Query ('/chat'):** This endpoint enables interaction with the **Meta Llama 3.1-8B Instruct** chatbot, which provides explanations based on CNN classification results. Clinicians can ask follow-up questions about tumor characteristics, diagnostic markers, and recommended treatments.

C.2. Chatbot Integration with Large Language Models (LLMs)

To enhance interpretability, the Flask API integrates a **Meta Llama 3.1-8B Instruct** chatbot that generates explanations of CNN predictions. The chatbot supports:

- **Interpreting CNN Outputs:** Upon receiving a classification result (e.g., Glioma), the chatbot explains the medical reasoning behind this classification, referencing tumor features commonly identified in MRI scans.
- **Providing Confidence Scores and Rationale:** The chatbot breaks down the CNN model's decision-making process, highlighting key radiological indicators such as tumor size, texture, and location.
- **Enabling Real-time Q&A with Clinicians:** Users can query additional details about differential diagnosis, histological subtypes, and **NCCN/ESMO** guideline recommendations. The chatbot leverages structured medical data to provide accurate, evidence-based responses.

By integrating the chatbot with CNN classification results, the system bridges the gap between AI-based tumor detection and clinical decision-making, providing enhanced interpretability for radiologists and oncologists.

Appendix D. Performance Evaluation and Ablation Study

D.1. Accuracy

Accuracy measures the proportion of correctly classified instances among all predictions. It provides an overall measure of correctness but may be misleading when dealing with imbalanced datasets.

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

where:

- **TP (True Positives)** = Correct tumor predictions
- **TN (True Negatives)** = Correct non-tumor predictions
- **FP (False Positives)** = Incorrect tumor predictions
- **FN (False Negatives)** = Missed tumor cases

For example, if a dataset contains 90% tumor cases, a model predicting all cases as tumors can still achieve 90% accuracy despite failing to identify non-tumor cases correctly.

D.2. Precision

Precision evaluates how many of the predicted tumor cases are actually tumors. This is particularly important in medical imaging, where false positives may lead to unnecessary treatments.

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

A high precision score means fewer false alarms, ensuring that most classified tumor cases are correct. However, focusing solely on precision may lead to missing actual tumor cases.

D.3. Recall (Sensitivity)

Recall, also known as sensitivity, measures the model's ability to detect actual tumor cases. A high recall ensures that most tumors are identified, reducing false negatives.

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

In brain tumor classification, high recall is crucial as missing a tumor diagnosis could lead to treatment delays. However, increasing recall may lower precision, necessitating a balance between the two.

D.4. F1 Score

The F1 Score is the harmonic mean of precision and recall, providing a balanced metric that is particularly useful when dataset classes are imbalanced.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

A high F1 score ensures the model effectively detects tumors while maintaining low false positives and false negatives, making it one of the most reliable performance metrics for medical applications.

D.5. Support

Support represents the number of true instances for each class in the dataset. It helps in evaluating the distribution of tumor types, ensuring that the model performs well across all categories (glioma, meningioma, pituitary tumor, and no tumor). While support itself does not have a formula, it is a key factor in interpreting precision, recall, and F1 scores.

D.6. Loss Function

The loss function quantifies how different the model's predictions are from the actual labels, guiding the learning process. For multi-class classification like brain tumor detection, categorical cross-entropy loss is commonly used:

$$\text{Loss} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{7}$$

where:

- y_i is the actual class label (1 for the correct class, 0 for others)
- $\hat{y}_i$ is the predicted probability for class i
- N is the number of classes

A lower loss value indicates better model performance. However, it must be evaluated alongside accuracy, precision, recall, and F1 score to ensure meaningful improvements in classification performance.

By analyzing these metrics together, deep learning models for brain tumor classification can be optimized for high accuracy, low misclassification rates, and reliable clinical application.