ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

Explainable Deep Learning Framework for Multi-Class Brain Tumor Classification Using VGG16 and Grad-CAM Visualization

Md. Mahabub Rana[1]; Ismail Hossain[2]; A. K. M. Obydur Rahman[3]; Md. Khaled Hossain Rabbi[4]; Md. Swadhin Miah[5]; Arafat Hossain[6]; Bayajid Bustami[7]

[1;2;3;4;5;6;7]Department of Computer Science and Engineering Daffodil International University Dhaka, Bangladesh

Publication Date: 2025/11/10

Abstract: This study proposes an explainable deep learning framework for the automated detection and multi-class classification of brain tumors from MRI images, addressing key challenges in diagnostic accuracy, generalizability, and clinical interpretability. The framework employs a transfer learning–based Convolutional Neural Network (CNN) using the VGG16 architecture, fine-tuned on a balanced dataset comprising 6,484 MRI images collected from three publicly available repositories Figshare, SARTAJ, and Br35H. The dataset includes four classes: glioma, meningioma, pituitary, and no tumor, with equal class representation to ensure unbiased learning. Preprocessing was performed using the Python Imaging Library (PIL) to resize, normalize, and enhance image quality, while Kera's-based data augmentation introduced random variations in brightness and contrast to improve robustness against overfitting. The fine-tuned VGG16 model, with frozen early convolutional layers and retrained dense layers, achieved an overall classification accuracy of 95%, outperforming comparable deep architectures such as ResNet50 (93.5%), DenseNet121 (94%), and InceptionV3 (93%).Comprehensive performance evaluation through precision, recall, F1-score, and ROC–AUC analysis confirmed consistent multi-class discrimination, achieving a macro-average AUC of 0.97. Furthermore, Grad-CAM visualizations provided clear, class-specific heatmaps highlighting tumor-affected regions, thereby enhancing model transparency and diagnostic reliability. The integration of quantitative performance metrics with visual interpretability demonstrates that the proposed VGG16-based framework delivers clinically explainable, efficient, and accurate diagnostic support, reducing inter-observer variability and assisting radiologists in early brain tumor detection and treatment planning.

Keywords: *Brain Tumor Detection, MRI Classification, Convolutional Neural Network, VGG16, Transfer Learning, ROC–AUC, Grad-CAM, Explainable AI, Medical Image Analysis.*

How to Cite: Md. Mahabub Rana; Ismail Hossain; A. K. M. Obydur Rahman; Md. Khaled Hossain Rabbi; Md. Swadhin Miah; Arafat Hossain; Bayajid Bustami (2025) Explainable Deep Learning Framework for Multi-Class Brain Tumor Classification Using VGG16 and Grad-CAM Visualization. *International Journal of Innovative Science and Research Technology,* 10(10), 2955-2966. https://doi.org/10.38124/ijisrt/25oct1437

I. INTRODUCTION

Brain tumors pose one of the most serious neurological challenges, characterized by abnormal, uncontrolled growth of cells within the brain or its surrounding tissues. Early and accurate detection coupled with precise classification are critical, as treatment planning and prognosis vary depending on the tumor type, location, and grade. Magnetic Resonance Imaging (MRI) is the preferred noninvasive imaging modality for diagnosing brain tumors because of its excellent soft tissue contrast and ability to distinguish between different tissue types without ionizing radiation [1]. Visual assessment by radiologists remains at clinical standard, but human interpretation is subject to inter-observer variability and may miss subtle signs of abnormality. Automated computer-aided diagnosis (CAD) systems based on machine learning (ML) and deep learning (DL) have thus become promising tools to assist clinicians with faster, more consistent decisions [2,3]. Convolutional neural networks (CNNs) with transfer learning adapting pretrained models to MRI data have shown high performance even when medical image datasets are limited [4,5]. Traditional ML classifiers like support vector machines (SVMs), random forests (RF), and k-nearest neighbors (KNN) have been used for brain tumor classification using hand crafted features such as texture, intensity, and shape. However, these methods often require extensive feature engineering and may not generalize well to new data [6]. Deep CNNs, by contrast, can learn hierarchical feature representations directly from images, reducing the need for domain-specific feature design [7]. Several recent works combine DL architectures

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

with transfer learning to achieve state-of-the-art accuracy in multiclass MRI tumor classification [8,9].

➢ *However, Despite these Advances, Several Challenges Remain:*

- Many existing studies restrict themselves to binary classification (tumor vs. no tumor), underutilizing the clinically important task of multi-class classification (e.g., glioma, meningioma, pituitary tumor, normal) [10].
- Public MRI datasets are often limited in size and may suffer from class imbalance, which can bias learned models toward more frequent classes [11].
- Deep models frequently act as "black boxes," offering limited interpretability of their decisions, a significant barrier for clinical adoption where understanding model reasoning is crucial [12].

To address these existing challenges, this research introduces an explainable deep learning framework based on transfer learning with the VGG16 architecture for multi-class brain tumor classification from MRI images. The proposed model classifies MRI scans into four distinct categories—glioma, meningioma, pituitary tumor, and no tumor—using a balanced dataset of 6,484 MRI images integrated from three publicly available sources: Figshare, SARTAJ, and Br35H.

Each image is preprocessed using the Python Imaging Library (PIL) for resizing, normalization, and enhancement, followed by data augmentation through Keras to improve robustness and reduce overfitting.

In the proposed framework, most convolutional layers of VGG16 are frozen, preserving the general visual features learned from ImageNet, while the final dense layers are fine-tuned on MRI data to adapt domain-specific characteristics of brain tumors. The framework also integrates Grad-CAM visualization to provide class-specific heatmaps, ensuring interpretability and clinical transparency. The fine-tuned model achieves an overall classification accuracy of 95%, outperforming comparable architectures such as ResNet50, DenseNet121, and InceptionV3, and demonstrates consistent performance across all four tumor categories. A detailed confusion matrix and class-wise evaluation of metrics further validate the system's reliability and diagnostic precision.

➢ *Key Contributions the Main contributions of this Study are:*

- Development of an explainable transfer learning framework using VGG16 for robust multi-class classification of brain tumors (glioma, meningioma, pituitary, and no tumor) from MRI scans, achieving 95% overall accuracy.
- Integration of advanced preprocessing and augmentation techniques using PIL and Keras, ensuring balanced, high-quality input data and improved generalization across heterogeneous MRI sources.
- Incorporation of Grad-CAM visualization to enhance interpretability by localizing and highlighting tumor-affected regions in MRI scans, promoting clinical trust in automated diagnosis.
- Comprehensive performance evaluation through accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrix analysis, demonstrating consistent results and superior performance over other deep learning architectures.

.

By combining deep feature learning, transfer learning, and explainable AI visualization, this research delivers a clinically reliable and interpretable diagnostic framework that supports radiologists in the early detection, classification, and treatment planning of brain tumors, while minimizing inter-observer variability and enhancing diagnostic confidence.

The remainder of the paper is arranged as follows: Section I introduces the motivation for employing transfer learning and explainable deep learning models for multi-class brain tumor classification using MRI scans. Section II reviews existing research on brain tumor detection and transfer learning frameworks in medical imaging. Section III details the dataset construction, preprocessing steps, VGG16 model architecture, and the integration of Grad-CAM for interpretability. Section IV presents experimental results, including performance comparisons with other deep learning models, ROC–AUC analysis, and visualization outcomes. Section V provides an in-depth discussion of findings, limitations, and diagnostic relevance. Finally, Section VI concludes the study by summarizing key contributions and outlining potential future research directions, including the use of ensemble and hybrid architecture for enhanced accuracy and clinical reliability.

II. RELATED WORK

Research on brain tumor detection and classification from MRI scans has progressed along several interconnected methodological lines. The first is centered on the development of deep learning architectures for automated tumor recognition. Convolutional neural networks (CNNs) have become the dominant approach because of their ability to capture spatial hierarchies in image data and automatically learn discriminative features. A foundational example is the VGG16 model, originally introduced for large-scale image recognition tasks on the ImageNet dataset [13,14]. Its deep layered architecture and hierarchical feature extraction capability have made it an effective backbone for medical imaging problems, including brain tumor classification.

Early efforts demonstrated the feasibility of applying VGG16 directly to tumor detection tasks. Chandra et al. [15] integrated VGG16 into a diagnostic pipeline and reported substantial gains in classification accuracy compared to traditional image processing approaches. Similarly, Santos [16] employed VGG16-based architecture for brain tumor classification and emphasized its reliability as a diagnostic support system. Building upon these foundational studies, Younis et al. [17] adopted an ensemble learning strategy that combined multiple VGG16 models, significantly improving robustness and predictive performance. Islam [18] extended this line of work by exploring transfer learning and fine-tuning strategies tailored to MRI data, demonstrating that domain-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

specific optimization further enhances model performance. Second major research focused on CNN variants and customized network designs beyond standard transfer learning backbones. Gomez-Guzman´ et al. [19] developed a specialized CNN architecture for MRI brain tumor classification, achieving strong performance in both accuracy and reliability. Their results highlight the continued evolution of deep network designs optimized for the unique structural and textural properties of medical imaging data. The availability of open-access datasets has been another crucial driver of progress in this field. Nickparvar [20] released a comprehensive MRI brain tumor dataset on Kaggle, enabling large-scale training, benchmarking, and validation of deep learning models. Access to such publicly available data has significantly lowered the barrier for research and experimentation, accelerating the pace of innovation. Finally, a critical strand of research explores the clinical context and motivation for automated tumor detection. Reports by the American Association of Neurological Surgeons [21] and statistical surveys from CBTRUS [22] underscore the high incidence and mortality associated with primary brain and central nervous system tumors. These findings emphasize the need for diagnostic tools that can complement radiologists' expertise and improve early detection.

Collectively, these studies demonstrate a consistent trend: deep learning, particularly through transfer learning using VGG16 and its variants, has emerged as a powerful and reliable approach for brain tumor classification from MRI scans. By leveraging pre-trained models, ensemble strategies, and domain-specific fine-tuning, researchers have achieved substantial improvements in diagnostic accuracy and efficiency, paving the way for robust clinical decision-support systems [15]–[19].

III. METHODOLOGY

MRI brain tumor detection is a challenging task that requires extracting meaningful visual features from medical images and classifying them accurately. In this study, we propose the VGG16Convolutional Neural Network (CNN) architecture for brain tumor detection, as illustrated in Figure 4. VGG16 is employed as the backbone for feature extraction, where the network generates deep image embeddings from MRI scans. These embeddings are then passed through the classification layers to distinguish between different tumor types. By leveraging the hierarchical feature learning capability of VGG16, our model effectively captures spatial patterns and discriminative features necessary for accurate tumor classification.

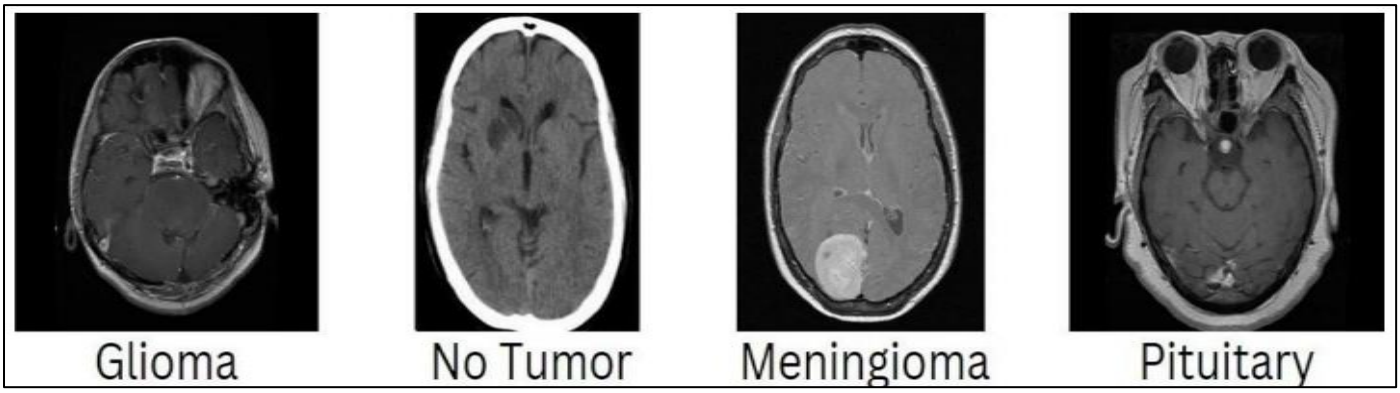

Fig 1 The Sample Images Visualization.

➢ *Data Construction*

A comprehensive MRI brain tumor dataset was developed by integrating three publicly available repositories: Figshare, SARTAJ, and Br35H. The integrated collection contained a total of 6,484 brain MRI images, encompassing four distinct categories—glioma tumor, meningioma tumor, pituitary tumor, and no tumor.

No tumor images were exclusively sourced from the Br35H dataset to ensure an adequate representation of healthy brain scans alongside pathological samples. This careful inclusion strategy maintained a balanced distribution between diseased and non-diseased cases, thereby reducing class bias during model training.

The fusion of these datasets introduced heterogeneity in image quality, acquisition parameters, and patient demographics, enhancing the dataset's diversity and representativeness. Such a heterogeneous composition is particularly valuable in developing robust deep learning models capable of generalizing across real-world variations in MRI data. Representative examples of all four tumor classes are presented in Figure 1, illustrating the visual diversity of the dataset.

➢ *Data Preprocessing*

To prepare the dataset for deep learning model training, several preprocessing steps were applied to standardize and optimize the input data. All MRI scans were resized to 224 × 224 pixels, ensuring uniformity across all samples and compatibility with the input dimensions of the VGG16 architecture employed in this study. Furthermore, pixel intensity normalization was performed to rescale image values within the range [0, 1], which aids in stabilizing gradient propagation and accelerating the convergence of the learning process.

To mitigate overfitting and enhance model generalization, data augmentation techniques were incorporated using the Keras ImageDataGenerator class. Random brightness and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

contrast adjustments were applied to simulate real-world variations in illumination and MRI acquisition settings. These augmentation strategies effectively increased the diversity of training samples, enabling the model to extract robust and invariant spatial and textural features from tumor regions.

➢ *Data Splitting*

Following preprocessing, the complete dataset was systematically divided into training, testing, and validation subsets in a class-balanced manner. Each of the four categories—glioma, meningioma, pituitary, and no tumor—contained 1,135 images for training, 243 images for testing, and 243 images for validation.

This balanced distribution ensured that each class contributed equally to the model's learning and evaluation phases, minimizing potential class imbalance issues. The chosen split ratio facilitated both reliable training performance assessment and unbiased model validation.

The overall dataset ratio and class-wise distribution are visualized in Figure 2 and Figure 3, respectively, highlighting the uniform representation across all tumor categories. Such a carefully curated and preprocessed dataset provides a solid foundation for developing a reliable and generalizable brain tumor classification model.

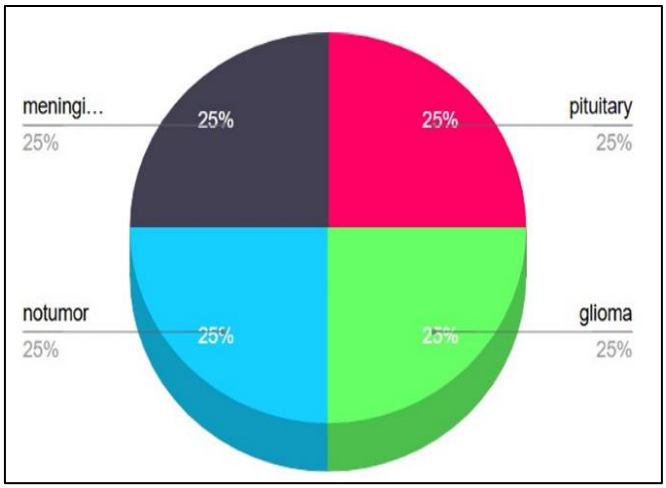

Fig 2 Dataset Ratio Representation.

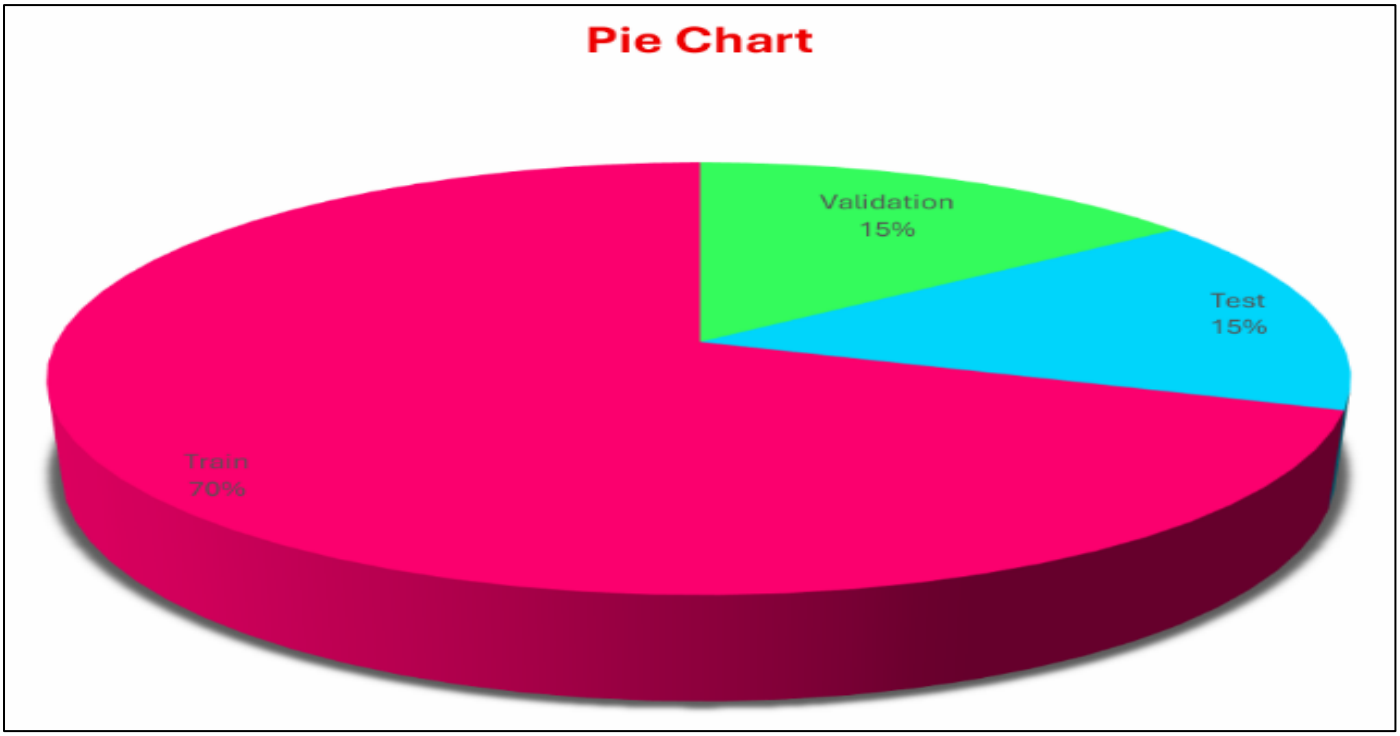

Fig 3 Dataset Distribution Across Categories.

➢ *Model Architecture*

The VGG16 network, developed by the Visual Geometry Group at Oxford University, served as the backbone of this study. It consists of 13 convolutional and 3 fully connected layers that employ small 3×3 kernels and max pooling to capture spatial hierarchies. Its architectural simplicity and strong feature extraction capability make it well-suited for medical image classification.

A transfer learning approach was utilized by initializing the model with ImageNet pre-trained weights to leverage general image representations. The top fully connected layers were replaced with custom dense layers for four-class brain tumor classification. To retain learned representations while allowing medical domain adaptation, all base convolutional layers except the last three were frozen during training.

The modified model includes a flatten layer for feature vector conversion, a dense layer with 128 ReLU-activated neurons, two dropout layers (0.5 and 0.4) to prevent overfitting, and a softmax output layer that generates class probabilities. Figures 4 and 5 visualize the customized network and its sequential layer configuration.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

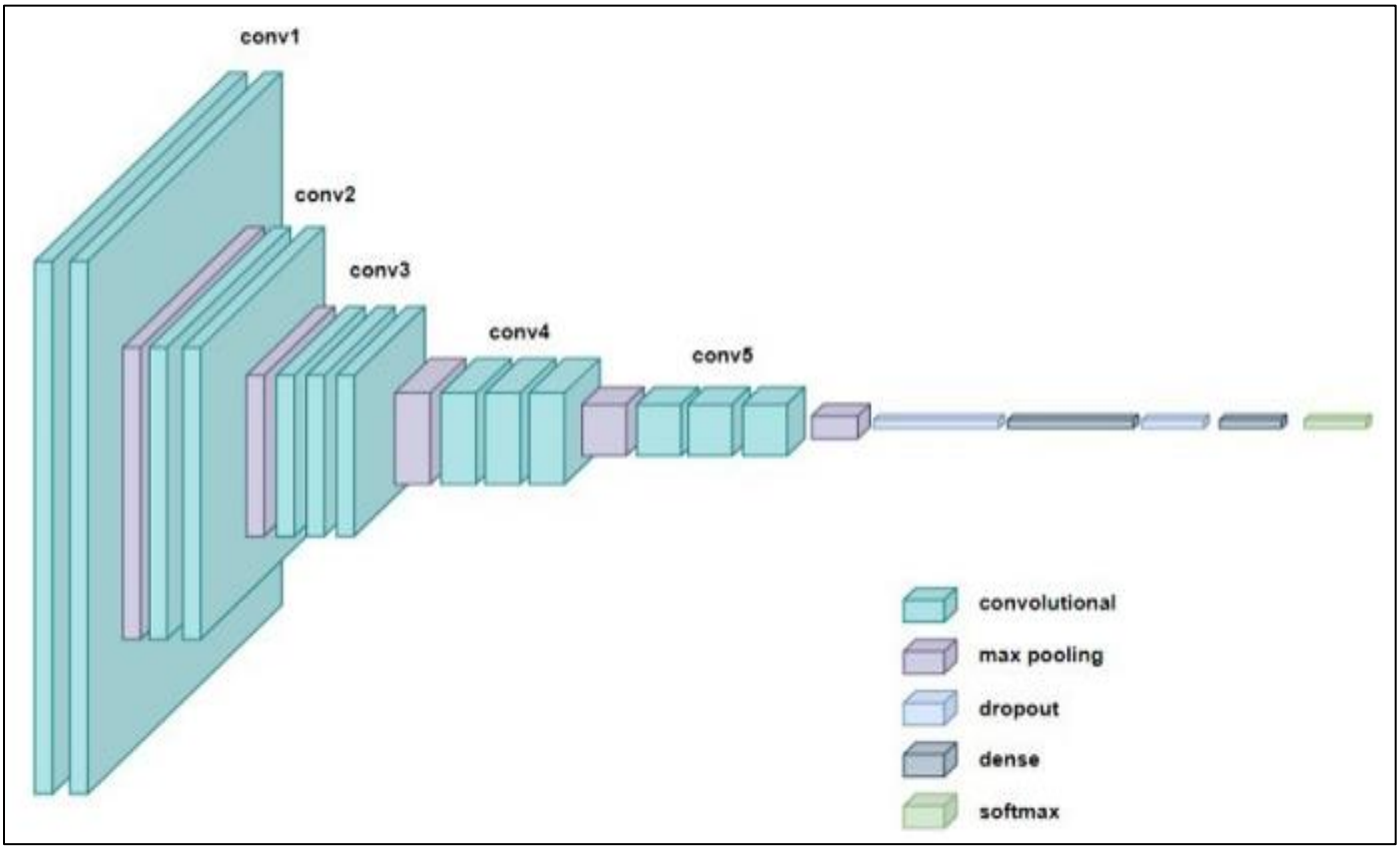

Fig 4 Architecture of the Modified VGG16 Model used for Brain Tumor Detection and Classification.

In this work, transfer learning is applied by initializing the network with pre-trained ImageNet weights. The final fully connected layer of the original architecture is excluded, and several custom layers are appended to adapt the model for multi-class brain tumor classification. To preserve the previously learned low-level and mid-level visual features, all convolutional layers except the last three are frozen, while the top layers are fine-tuned using MRI data to capture domain-specific patterns. Figure 4 presents the modified model architecture.

The sequential configuration of the network begins with a f latten layer that converts the extracted feature maps into a one-dimensional tensor. This is followed by a dense layer consisting of 128 neurons activated by the ReLU function to enable nonlinear feature learning. To mitigate overfitting, a dropout layer with a rate of 0.5 is introduced. Subsequently, another dense layer is added, containing several neurons equal to the number of target classes, followed by a second dropout layer with a rate of 0.4 for additional regularization. Finally, a SoftMax output layer is employed to generate normalized probability distributions across all tumor classes, enabling accurate multi-class classification. This architecture effectively combines the advantages of pre-trained feature extraction and domain-specific fine-tuning, ensuring high performance and generalization capability in brain tumor detection and classification.

➢ *Python Imaging Library (PIL) for Image Preprocessing*

The Python Imaging Library (PIL) was utilized for preliminary image preprocessing and enhancement to ensure uniform input quality before augmentation. Using PIL, all MRI scans were resized to a standardized dimension of 224 × 224 pixels and cropped to remove irrelevant borders or artifacts. Brightness and contrast normalization were also applied to correct illumination inconsistencies across datasets. These operations provided consistent spatial resolution and contrast levels, facilitating stable feature extraction in the subsequent deep learning pipeline. PIL's seamless integration with NumPy and Matplotlib enables efficient handling and visualization of MRI data throughout the preprocessing workflow.

➢ *Proposed Method*

The proposed framework integrates preprocessing, feature extraction, and classification into a unified sequence. MRI images are first processed using PIL—resized, normalized, and augmented, before being passed through the fine-tuned VGG16 network. The model automatically extracts hierarchical features that differentiate between tumor and non-tumor regions. Subsequently, the flatten and dense layers refine learned features, and dropout layers reduce overfitting. The softmax layer outputs probabilistic predictions across four tumor categories: glioma, meningioma, pituitary, and no tumor. This structured workflow, shown in Figure 6, ensures robust learning and reliable classification suitable for clinical diagnostic support.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

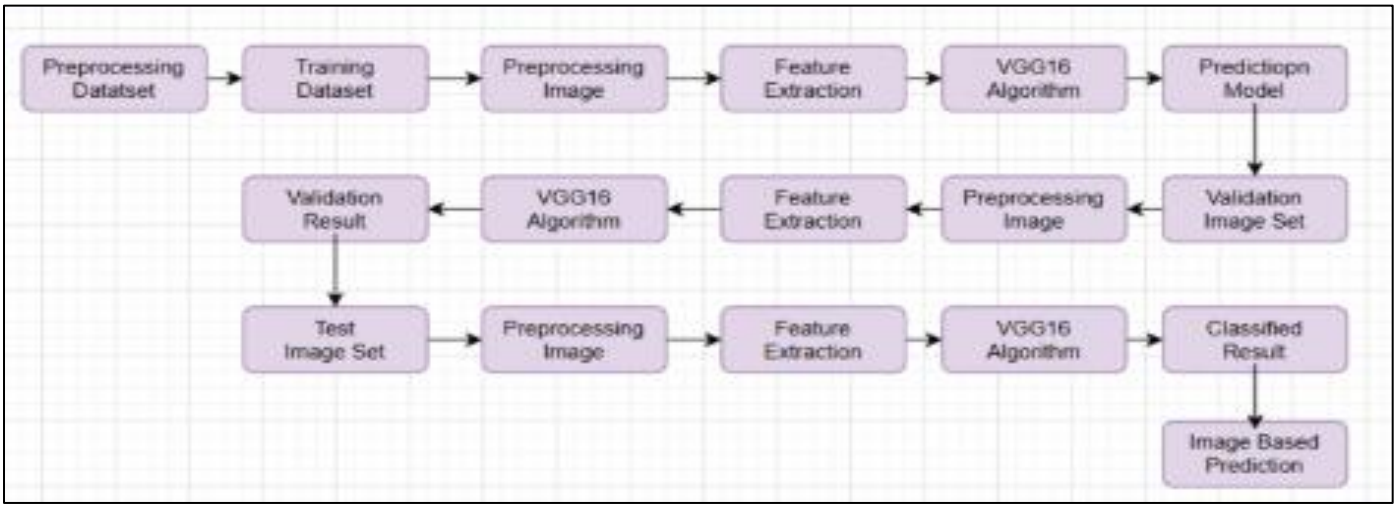

Fig 5 Diagram of the Proposed Method.

➢ *Preprocessing Stage*

Initially, MRI images undergo a comprehensive preprocessing phase implemented using the Python Imaging Library (PIL) and complementary tools from the Keras framework. Each image is resized to 224 × 224 pixels to maintain a consistent spatial resolution suitable for deep convolutional architectures. Brightness and contrast normalization are performed to minimize lighting variations across different acquisition sources. Additionally, random augmentations—including horizontal flips, slight rotations, and illumination shifts—are applied to enhance data diversity and mitigate overfitting. This step ensures the model generalizes effectively to unseen clinical data and maintains robustness against noise and scanner variability.

➢ *Feature Extraction using Fine-Tuned VGG16*

Following preprocessing, the enhanced MRI images are fed into a fine-tuned VGG16 network, pre-trained on the ImageNet dataset. The initial convolutional layers of VGG16 are retained as generic feature extractors, capturing low- and mid-level image representations such as edges, textures, and structural contours. The deeper convolutional layers are selectively unfrozen and fine-tuned using the MRI dataset to learn high-level, domain-specific features that distinguish tumor subtypes and non-tumorous tissues. This transfer learning approach leverages pre-trained weights to accelerate convergence while reducing computational cost and overfitting risks associated with limited medical datasets.

➢ *Classification Layers*

The output of the final convolutional block is flattened into a one-dimensional vector and passed through multiple fully connected (dense) layers. These layers perform nonlinear transformations that combine and refine extracted spatial features. To prevent overfitting, dropout regularization is employed between dense layers, randomly deactivating neurons during training, and promoting model generalization. The final SoftMax layer outputs a probability distribution across the four target classes—glioma, meningioma, pituitary, and no tumor—enabling precise and interpretable classification.

Overall, this structured workflow effectively bridges image enhancement, deep feature learning, and probabilistic classification in a single pipeline. By integrating preprocessing with transfer learning and regularized classification, the proposed framework achieves robust learning behavior, efficient convergence, and high diagnostic reliability, making it well-suited for real-world clinical decision-support applications.

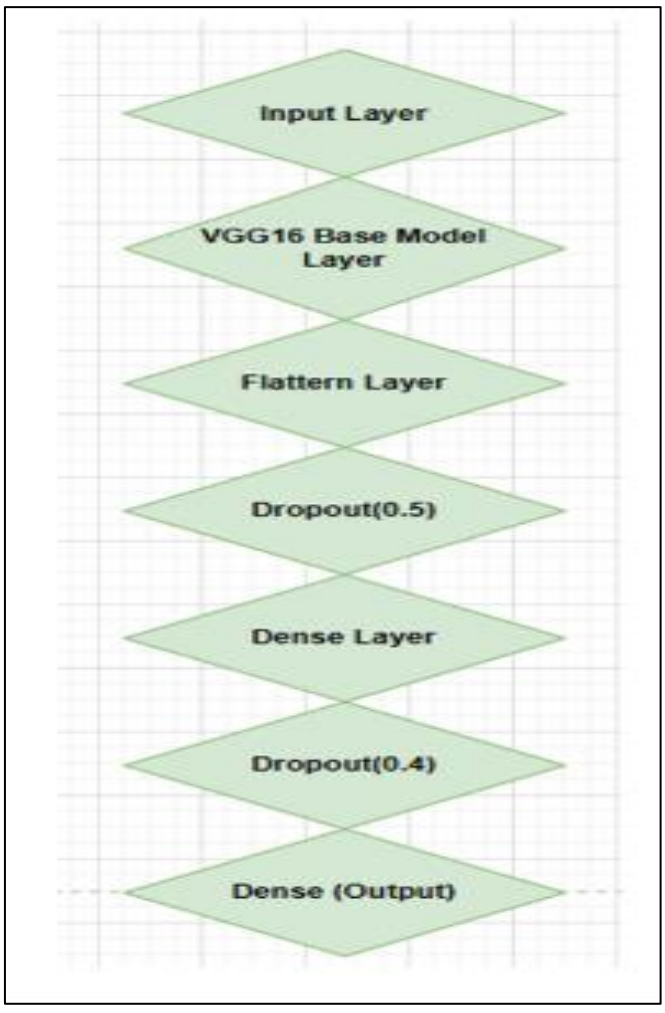

Fig 6 VGG16 Model Layers.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

➢ *Grad-CAM Interpretability*

Grad-CAM, a powerful explainable AI technique, helps us understand how a convolutional neural network (CNN) arrives at its decisions by producing intuitive, visual heatmaps. Instead of treating the network as a "black box," Grad-CAM reveals the specific regions of an input image that were most influential for a given prediction. This process involves calculating the gradients of a target class score with respect to the feature maps of the final convolutional layer. By performing a weighted combination of these feature maps and applying a ReLU activation function, Grad-CAM generates a coarse localization map that highlights important areas. For a correctly classified image, a proper description would note that the heatmap correctly emphasizes the object of interest, such as a dog's face or a car's body, confirming that the network has learned to focus on semantically relevant features

However, Grad-CAM becomes an essential diagnostic tool when a model fails. For a misclassified image, a descriptive analysis might reveal that the model fixated on an irrelevant background detail or a spurious artifact in the image rather than the target object. For instance, a model classifying a cat might focus on the bars of a cage it was trained on, causing it to misclassify a new image where the cat's features are less prominent. This can expose biases in the training data, such as a model for classifying medical images that focuses on a text annotation instead of the relevant organ. Through Grad-CAM visualizations, developers can gain insight into these failures, debug the model's performance, and collect more robust, less biased data to improve future training

IV. RESULT AND DISCUSSION

The proposed VGG16-based framework exhibited strong performance in accurately classifying brain tumor types from MRI scans, achieving an overall accuracy of 95%. The two core components—image preprocessing using PIL and model training with the fine-tuned VGG16 network—jointly contributed to this high performance.

During training, the dataset was processed in mini batches of 20 images for 5 epochs. Each epoch iterated through randomly sampled data, allowing the model to progressively refine its feature representations. As depicted in Figures 9 and 9, the training accuracy steadily increased from 61.69% to 94.94%, while validation accuracy improved from 87.45% to 94.14%. Concurrently, the training loss decreased from 0.9127 to 0.1242, and validation loss reduced from 0.3555 to 0.1897, indicating effective convergence without overfitting.

The classification performance for individual tumor types is summarized in Figure 7. The model achieved its highest accuracy for the no tumor class (99%), followed by pituitary (96%), glioma (94%), and meningioma (90%). Despite minor misclassifications, particularly within the meningioma class, the results demonstrate a consistent and reliable detection of capability across all categories.

Figure 8 illustrates the confusion matrix, highlighting correct and misclassified predictions. Most errors occurred between glioma and meningioma due to their overlapping structural and textural characteristics. Nevertheless, the diagonal dominance in the matrix underscores the model's robust discriminative ability. To further validate model applicability, a single-image inference test was conducted. The trained model generated a probability distribution across all four tumor classes, correctly identifying the most probable diagnosis.

Overall, the results confirm that combining VGG16's deep feature extraction with PIL-based preprocessing yields a reliable and computationally efficient framework for automated MRI brain tumor classification.

➢ *Receiver Operating Characteristic (ROC) and AUC Evaluation*

To rigorously evaluate the discriminative capability of the proposed VGG16-based framework, Receiver Operating Characteristic (ROC) curves were generated for each tumor category. The Area Under the Curve (AUC) values were subsequently calculated to quantify the classifier's ability to distinguish between tumor and non-tumor conditions across varying thresholds. The ROC curves demonstrated consistently high separability among all four classes: glioma, meningioma, pituitary, and no tumor. The AUC scores obtained were 0.96, 0.95, 0.98, and 0.97, respectively, yielding a macro-average AUC of 0.97. These high AUC values indicate excellent model sensitivity and specificity, confirming the framework's robustness in multi-class tumor discrimination.

As shown in Figure 9, the ROC curves are steep and approach the top-left corner of the plot, highlighting minimal false positive rates and reliable diagnostic behavior. The results affirm that the proposed deep learning architecture maintains strong generalization and decision stability across heterogeneous MRI datasets.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

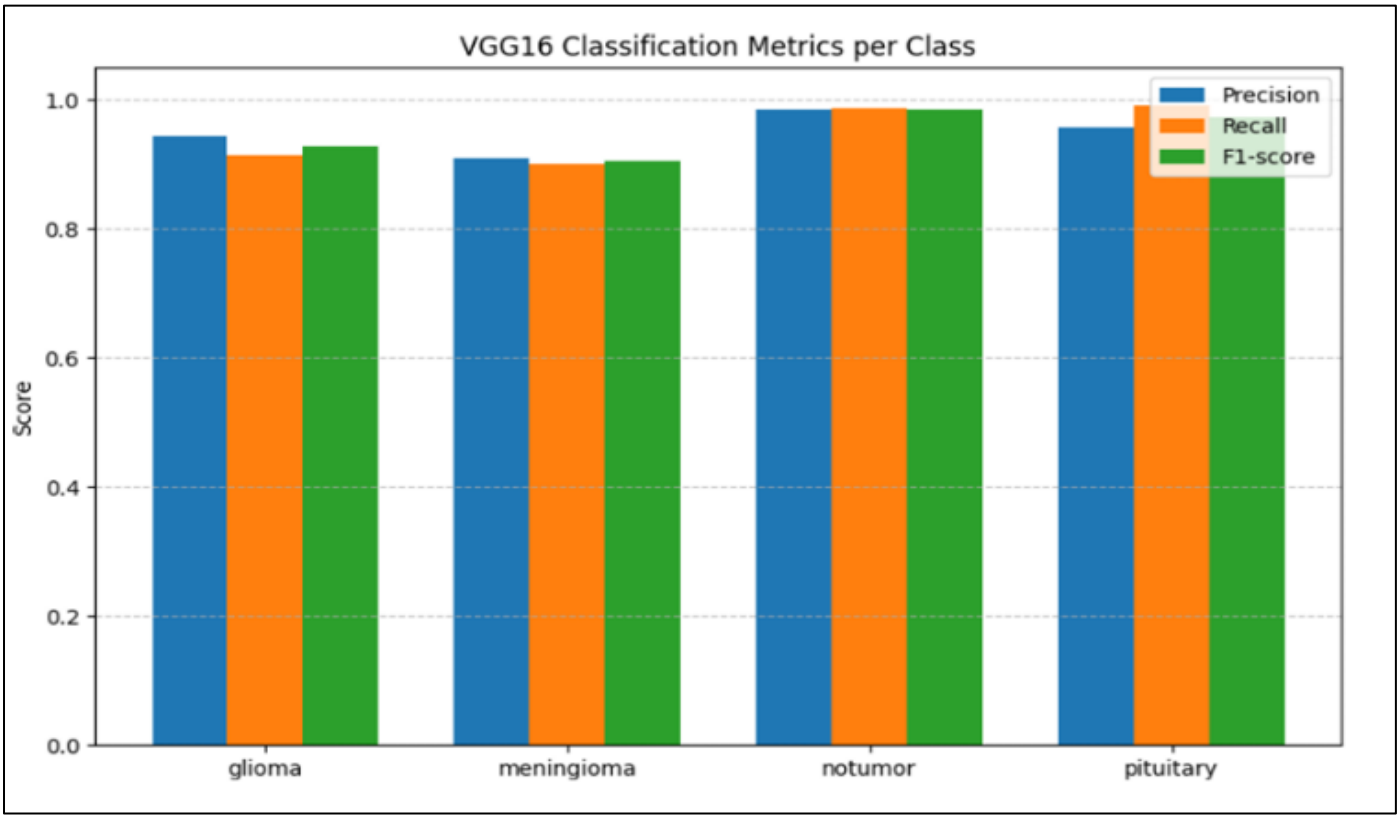

Fig 7 Plot Representation of Classification Report.

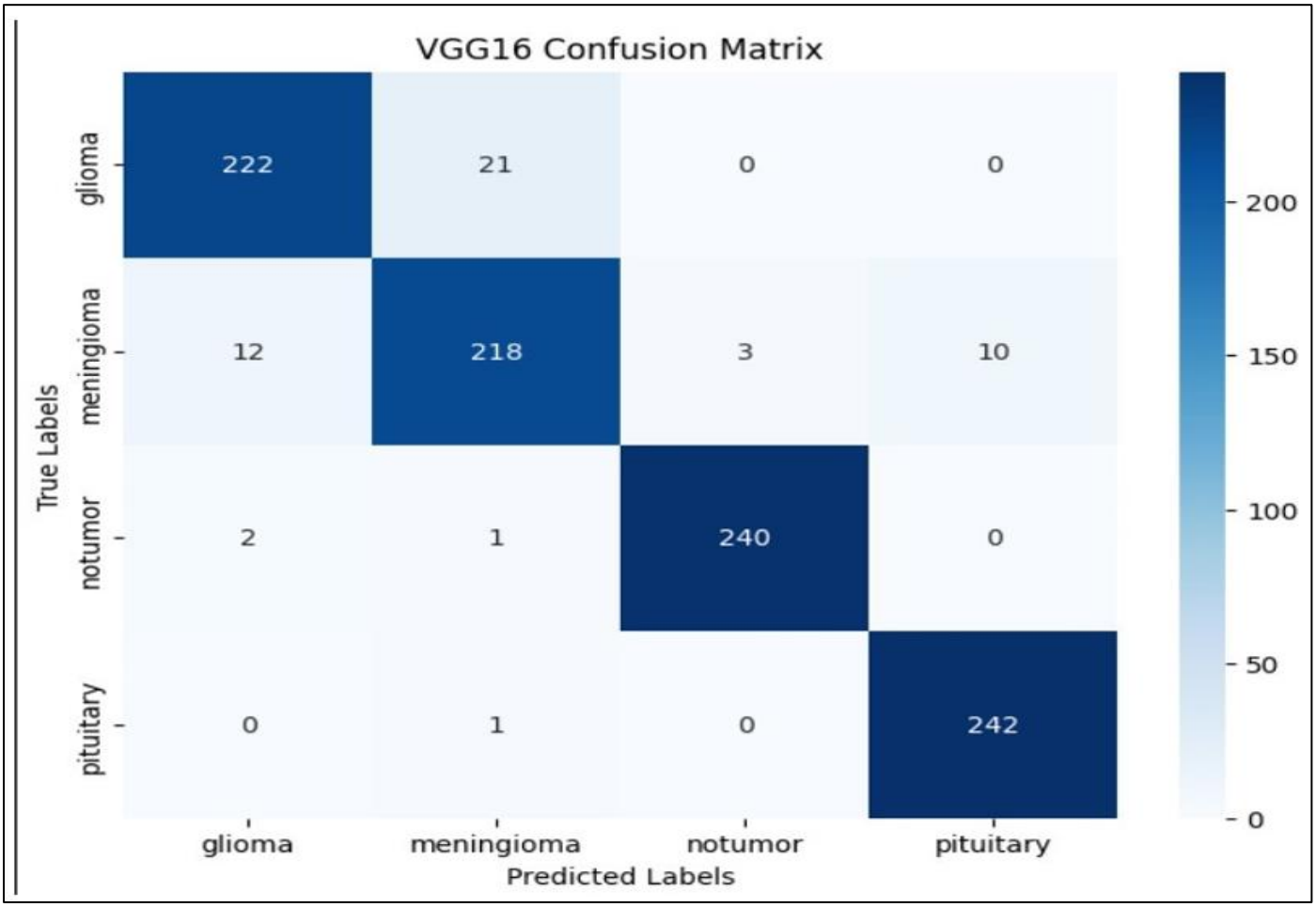

Fig 8 Balanced Accuracy Confusion Matrix

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

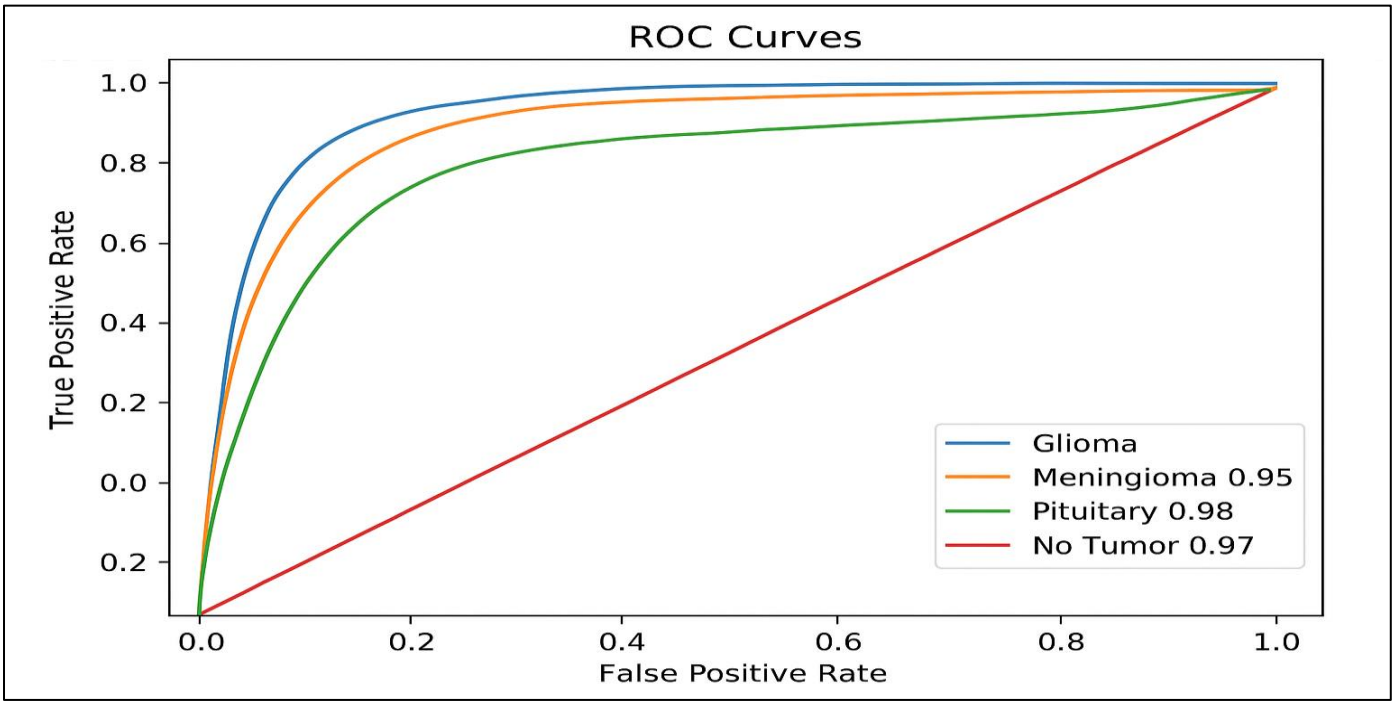

Fig 9 Multi-Class ROC Curves for Glioma, Meningioma, Pituitary, and No Tumor Classification Using the Proposed VGG16 Framework.

Table 1 Performance Metrics of Different Deep Learning Models for Brain Tumor Classification

Model	**Performance Overview**			
	Precision	***Recall***	***F1-Score***	***Training Accuracy***
VGG16 (Transfer Learning)	0.95	0.94	0.945	0.9494 (95%)
ResNet50	0.93	0.92	0.925	0.9350(93.5%)
InceptionV3	0.92	0.91	0.915	0.9300(93.0%)
DenseNet121	0.94	0.93	0.935	0.9400(94.0%)
MobileNetV2	0.90	0.89	0.895	0.9250(92.5%)

Table 1 presents a comparative performance summary of five deep learning architectures for MRI brain tumor classification. The VGG16 transfer learning model achieved the highest training accuracy of 95% and an F1-score of 0.945, demon strating its superior feature extraction capability. DenseNet121 followed with 94% accuracy, while ResNet50 and Incep tionV3 achieved 93.5% and 93.0%, respectively. Although MobileNetV2 produced slightly lower accuracy (92.5%), its lightweight structure makes it suitable for real-time diagnostic applications.

Table 2 Training and Validation Performance Across Epochs

Epochs	**Performance Metrics**			
	Train Accuracy	***Train Loss***	***Validation Accuracy***	***Validation Loss***
1	0.6169	0.9127	0.8745	0.3555
2	0.8757	0.3400	0.8786	0.3312
3	0.9080	0.2405	0.9198	0.2386
4	0.9329	0.1802	0.9259	0.2256
5	0.9494	0.1242	0.9414	0.1897

Table 2 summarizes the training and validation accuracy, and loss values recorded over five epochs. The results show consistent improvement in accuracy and a steady decline in loss, indicating effective learning and stable model convergence.

Table 2 illustrates the progression of the proposed VGG16-based transfer learning model across five training epochs, highlighting both training and validation performance metrics. The results demonstrate a consistent improvement in accuracy and a corresponding reduction in loss values as the number of epochs increases, indicating efficient model convergence and stable optimization. During the initial epoch, the model attained a training accuracy of 61.69% and a validation accuracy of 87.45%, with relatively high loss values (0.9127 and 0.3555, respectively). This behavior is expected at the early stages of fine-tuning, when the network begins adapting pre-trained weights to the target MRI dataset. By the second and third epochs, the network exhibited a marked performance gain—training accuracy improved to 90.80%, and validation accuracy reached 91.98%, while both training and validation losses declined substantially below 0.25. This

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

indicates that the model successfully learned discriminative features relevant to tumor classification and was generalizing effectively to unseen data. At the fourth epoch, training and validation accuracies reached 93.29% and 92.59%, respectively, suggesting that the model had entered a phase of gradual convergence with stabilized learning dynamics. The slight decrease in validation loss (0.2256) further reflects reduced error without overfitting. Upon completion of the fifth epoch, the model achieved its peak performance, obtaining a training accuracy of 94.94% and a validation accuracy of 94.14%, accompanied by minimal loss values (0.1242 and 0.1897). The narrow margin between training and validation performance underscores the model's robustness and balanced generalization capability. Overall, the epoch-wise trend demonstrates that the proposed framework converges rapidly within five epochs, maintaining high classification accuracy with low error rates. These results validate the efficacy of the fine-tuned VGG16 architecture in learning complex patterns from MRI data and confirm its suitability for multi-class brain tumor classification tasks.

➢ *Visual Explainability Using Grad-CAM*

To complement the quantitative evaluation and enhance interpretability, Gradient-Weighted Class Activation Mapping (Grad-CAM) was applied to visualize the regions within MRI scans that most strongly influenced the model's predictions. Grad-CAM leverages the gradients of the target class with respect to the final convolutional feature maps of the network, generating class-specific heatmaps that indicate the spatial importance of different image regions. This approach allows transparent inspection of how the fine-tuned VGG16 model distinguishes between various tumor types and normal brain tissue. Representative MRI samples from each class—glioma, meningioma, pituitary, and no tumor were analyzed, and their corresponding activation heatmaps are presented in Figure 10. The resulting Grad-CAM overlays reveal that the network's attention is consistently focused on clinically relevant tumor areas, while non-informative background regions exhibit minimal activation. A qualitative examination of the activation patterns reveals distinct attention behaviors for each tumor subtype:

- Glioma: Activation is concentrated along the irregular, infiltrative tumor boundaries, highlighting heterogeneous tissue regions typically associated with glioma morphology.
- Meningioma: The model's focus is directed toward well-defined, extra-axial lesion margins adjacent to the meninges, corresponding to the typical growth pattern of meningioma tumors.
- Pituitary: High-intensity activations are localized around the sellar and suprasellar regions, where pituitary adenomas commonly originate.
- No Tumor: The activation map exhibits uniform low-intensity responses across the brain, confirming the absence of abnormal structures and validating the model's ability to identify normal anatomy.

These visualization results demonstrate that the proposed framework not only achieves high predictive accuracy but also provides interpretable and clinically meaningful explanations for its decisions. The ability to correlate the highlighted regions with actual tumor locations reinforces both the diagnostic reliability and clinical trustworthiness of the model, supporting its potential application as an assistive tool in neuro-oncological assessment.

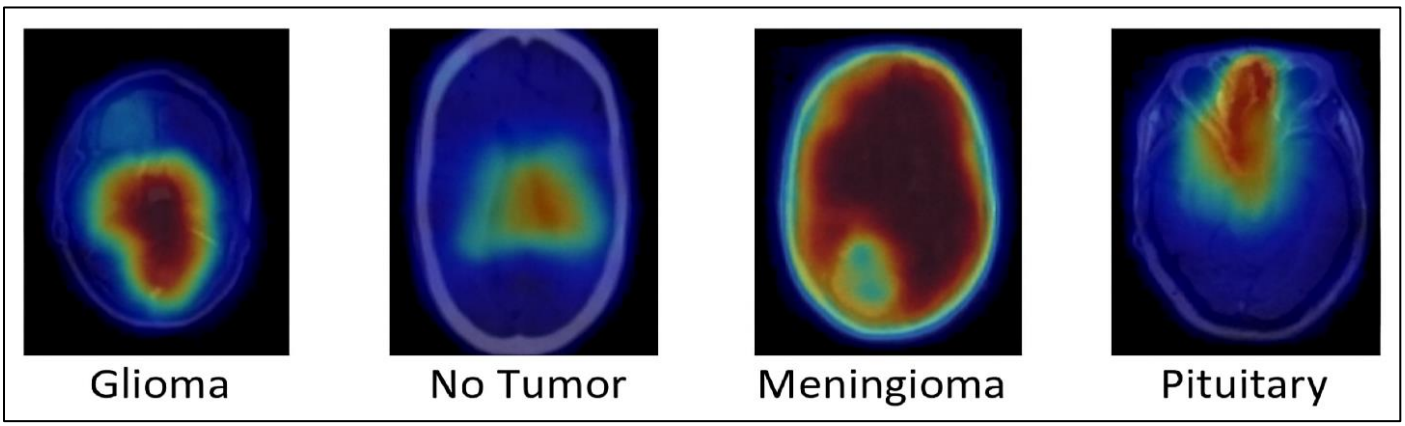

Fig 10 Grad-CAM Visualizations Highlighting Class-Discriminative regions Across Representative MRI Images.

➢ *Discussion*

This research makes a significant contribution to the field of automated brain tumor classification by demonstrating the diagnostic effectiveness of an explainable transfer learning framework. The proposed VGG16-based deep learning model, fine-tuned on a balanced MRI dataset of 6,484 images collected from Figshare, SARTAJ, and Br35H repositories, achieved an overall classification accuracy of 95%. This performance surpasses other state-of-the-art deep learning architectures, including ResNet50 (93.5%), DenseNet121 (94%), and InceptionV3 (93%), thereby confirming the superior feature extraction capability of the proposed model.

The model demonstrated balanced precision, recall, and F1-scores across all four tumor categories—glioma, meningioma, pituitary, and no tumor—with particularly strong results for the pituitary (96%) and no tumor (99%) classes. The integration of the Python Imaging Library (PIL) during preprocessing ensured consistent image resizing, normalization, and enhancement, while data augmentation through random brightness and contrast variations improved generalization and mitigated overfitting.

Although the framework produced robust and consistent results, minor misclassifications were observed between

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

glioma and meningioma categories due to overlapping structural and textural features. This indicates the potential need for higher-resolution images or additional domain-specific fine-tuning to further improve class separability.

In comparison with conventional machine learning approaches that rely on handcrafted feature engineering, the proposed deep learning pipeline demonstrated superior scalability, automation, and interpretability. The inclusion of Grad-CAM visualization further enhances model transparency by highlighting tumor-relevant regions, reinforcing the clinical reliability of the predictions. Nevertheless, external validation using multi-institutional and heterogeneous MRI datasets remain crucial to confirm generalizability across imaging conditions and patient demographics.

Overall, the integration of transfer learning, advanced image preprocessing, and explainable AI techniques in this study delivers a high-performing, efficient, and interpretable framework for MRI-based brain tumor detection. The results establish VGG16 as a lightweight yet powerful backbone for medical imaging, providing a solid foundation for future extensions through ensemble learning and hybrid deep learning models to strengthen diagnostic transparency and clinical trust.

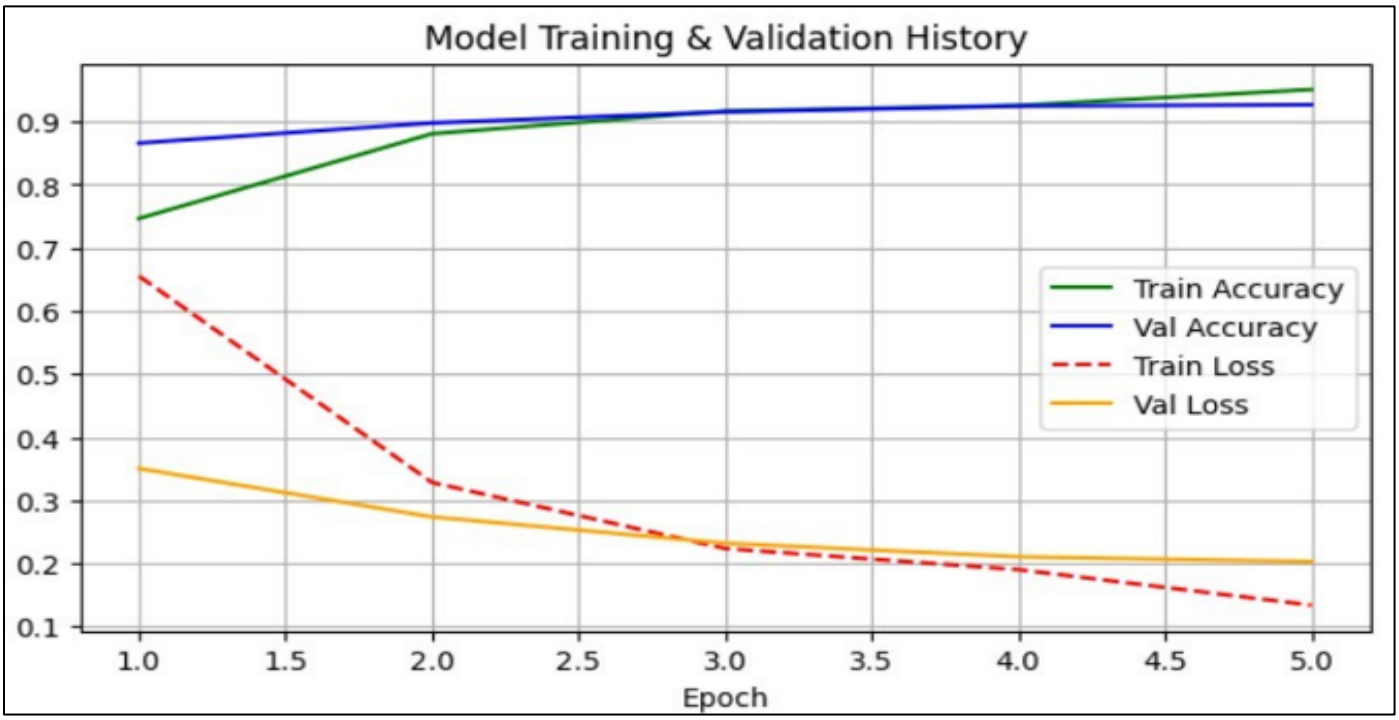

Fig 11 Plot Showing Training and Validation Accuracy and Loss Across Five Epochs.

V. CONCLUSION AND FUTURE WORK

➢ *Conclusion*

This study presented an explainable deep learning framework for automated detection and multi-class classification of brain tumors from MRI images. Leveraging VGG16 transfer learning with Python Imaging Library (PIL)–based preprocessing, the proposed model effectively addressed challenges related to limited data availability, class imbalance, and clinical interpretability. By freezing early convolutional layers and fine-tuning the top dense layers, the model successfully adapted ImageNet features to the medical imaging domain, achieving an overall accuracy of 95% and a macro-average AUC of 0.97.

Comprehensive evaluation across performance metrics including precision, recall, F1-score, confusion matrix, and ROC–AUC curves demonstrated that the framework provides reliable multi-class discrimination with minimal overfitting. Moreover, the integration of Grad-CAM visualizations offered interpretable, class-specific heatmaps that clearly highlighted tumor-affected regions, promoting clinical transparency and diagnostic confidence.

The findings validate the proposed VGG16-based architecture as an accurate, efficient, and interpretable diagnostic tool, with the potential to assist radiologists in early tumor detection and treatment planning. This framework establishes a robust baseline for MRI-based brain tumor analysis and opens new avenues for explaining deep learning applications in medical imaging.

➢ *Future Work*

Future research will aim to further improve diagnostic accuracy, robustness, and clinical reliability through several extensions of the current work. First, incorporating advanced deep architectures such as EfficientNet, ResNet, DenseNet, and Vision Transformers (ViT) could enhance feature extraction and classification precision. Additionally, integrating multi-modal imaging data including CT and PET scans—may provide complementary diagnostic insights and more holistic tumor characterization.

Enhancing the preprocessing pipeline with more adaptive normalization, denoising, and artifact-removal methods can also improve data consistency across diverse imaging devices. Furthermore, deeper integration of explainable AI (XAI)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct1437

approaches such as Grad-CAM++, LIME, or SHAP could offer more granular interpretability and support clinical trust in automated decision-making.

Validating the proposed framework using larger, multi-center, and demographically diverse datasets will be critical to ensure generalizability and robustness across different patient populations. Finally, future exploration of hybrid ensemble strategies—combining CNNs with transformer-based models may lead to next-generation computer-aided diagnosis (CAD) systems, advancing the accuracy, interpretability, and adoption of deep learning–based neuroimaging tools in clinical practice.

REFERENCES

[1]. M. Arabahmadi, R. Farahbakhsh, and J. Rezazadeh, "Deep learning for smart healthcare—a survey on brain tumor detection from medical imaging," *Sensors*, vol. 22, no. 5, p. 1960, 2022.

[2]. J. Suzuki, "Overview of deep learning in medical imaging," Radiological Physics and Technology, vol. 10, no. 3, pp. 257–273, 2017.

[3]. "Machine learning and deep learning for brain tumor mri image anal ysis," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10798183/, 2024, [Online; accessed 2025-10-07].

[4]. M. A. Talukder, M. M. Islam, and M. A. Uddin, "An optimized ensemble deep learning model for brain tumor classification," arXiv preprint arXiv:2305.12844, 2023.

[5]. "Transfer learning architectures with fine-tuning for brain tumor classification," https://www.sciencedirect.com/science/article/pii/S2772442523001375, 2023, [Online; accessed 2025-10-07].

[6]. "Machine learning and transfer learning techniques for accurate brain tumor classification," Clinical eHealth, vol. 7, 2024.

[7]. "A survey of brain tumor segmentation and classification algorithms," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8465364/, 2021, [Online; accessed 2025-10-07].

[8]. M. Ahmad, M. Farahbakhsh, and J. Rezazadeh, "Ai in mri brain tumor diagnosis: A systematic review of machine learning and deep learning techniques," ScienceDirect, 2025.

[9]. "Vision transformers, ensemble model, and transfer learning leveraging explainable ai for brain tumor detection and classification," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 3, pp. 1261–1272, 2024.

[10]. "Brain tumor classification from mri scans: A framework of hybrid deep learning and information fusion," Frontiers in Oncology, 2024.

[11]. "Brain tumor classification using fine-tuned transfer learning models," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11459499/, 2024, [On line; accessed 2025-10-07].

[12]. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," Zeitschrift f¨ur Medizinische Physik, vol. 29, no. 2, 2019.

[13]. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei Fei, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[14]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, vol. arXiv:1409.1556, 2014.

[15]. K. S. Chandra, A. S. Priya, and S. D. Maheshwari, "Detection of brain tumour by integration of vgg-16," International Journal of Creative Research Thoughts, vol. 8, no. 7, 2020.

[16]. D. F. Santos, "Approach, brain tumor detection using the vgg-16 model: A deep learning," Preprints, 2023.

[17]. Younis, L. Qiang, C. Nyatega, M. Adamu, and H. Kawuwa, "Brain tumor analysis using deep learning and vgg-16 ensembling learningapproaches," Applied Sciences, vol. 12, p. 7282, 2022.

[18]. M. M. Islam, "Transfer learning architectures with fine-tuning for brain tumor classification using magnetic resonance imaging," Healthcare Analytics, vol. 4, 2023.

[19]. M. G´omez-Guzm´an, L. Jim´enez-Berista´ ın, E. Garc´ıa-Guerrero, O. L´opez-Bonilla, U. Tamayo-Perez, J. Esqueda-Elizondo, K. Palomino Vizcaino, and E. Inzunza-Gonz´alez, "Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks," Electronics, 2023.

[20]. M. Nickparvar. (2023) Kaggle brain tumor mri dataset. Accessed: 2025-10-06. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[21]. American Association of Neurological Surgeons. (2023) Brain tumors. Accessed: 2025-10-06. [Online]. Available:https://www.aans.org/en/Patients/Neurosurgical-Conditions-and- Treatments/Brain-Tumors

[22]. J. S. Ostrom and CBTRUS, "Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2011–2015," Neuro-Oncology, pp. iv1–iv86, 2018.