Date available online: 09-01-2026
Vol. 3, Issue 2 (July – December 2025)

DOI: https://doi.org/10.24312/ucp-jeit.03.02.573

Multi-Class Brain Tumor Detection Using Transfer Learning and Interpretable Deep Models

Muhammad Asif Feroz[1], Anam Safdar Awan[2], Fareeha Batool[3], Narges Shahbaz[4], Anam Murtaza[5], Kamran Ali[6]

[1]Department of Computer Science and IT, Superior University, Lahore, 54000, Pakistan (e-mail: asif.janjua005@gmail.com)
[2]Department of Computer Science and IT, Superior University, Lahore, 54000, Pakistan (e-mail: anamawan416@gmail.com)
[3]Department of CS & IT, Superior University, Sargodha Campus, Pakistan (e-mail: fareehabaloch29@gmail.com)
[4]Department of Arts and Humanities, Superior University, Lahore, Pakistan (e-mail: nargesshahbaz20137@gmail.com)
[5]Department of Biology, Superior University, Sargodha Campus, Pakistan (e-mail: anammurtazashah434@gmail.com)
[6]Department of CS & IT, University of Sargodha, Sargodha, Pakistan (e-mail: kamranali3037414518@gmail.com)

Corresponding author: Muhammad Asif Feroz (e-mail: asif.janjua005@gmail.com).

ABSTRACT

Accurate brain tumor detection remains critical yet challenging due to diagnostic complexity and variability in MRI interpretation. This study proposes a deep learning approach for automated multi-class brain tumor classification using transfer learning (TL). Three pre-trained CNN models, ResNet50, InceptionV3, and VGG16, were adapted and evaluated on a curated MRI dataset of 7,000+ images. Preprocessing, feature extraction, fine-tuning, and integration of Explainable AI (Grad-CAM, LIME, SHAP) ensured robust and interpretable results. ResNet50 achieved the highest performance with 98% accuracy, 0.92 F1-score, and 0.96 AUC, outperforming the other models across all metrics, with strong convergence and minimal misclassification. ResNet50's architecture enabled deeper feature learning and improved generalization. Explainable AI visualizations confirmed model focus on tumor-relevant MRI regions, enhancing clinical interpretability. The findings position ResNet50 as an effective and explainable solution for MRI-based brain tumor classification, suitable for future real-world deployment and further expansion to mobile and multi-center applications.

INDEX TERMS: Brain Tumor Detection, Deep Learning, ResNet50, MRI Classification, CNN, Medical Imaging, Binary Classification, Tumor Diagnosis.

I. INTRODUCTION

Brain tumors are among the most critical neurological disorders, characterized by aberrant development of cells inside or around the brain that perturb normal brain function [1]. Depending on their nature, brain tumors are often classed into *benign* (non-cancerous and slow-growing) and *malignant* (cancerous and aggressive) [2]. According to the International Agency for Research on Cancer (IARC), more than 126,000 new brain tumor cases are diagnosed annually worldwide, with over 97,000 deaths attributed to the disease each year. The World Health Organization (WHO) further projects a 5% annual increase in brain tumor cases globally, making early detection and effective treatment increasingly vital [3; 4].

The early and correct diagnosis of brain tumors plays a vital role in enhancing patient outcomes, reducing mortality rates, and planning personalized treatment strategies [5]. Magnetic Resonance Imaging (MRI) has evolved as a key imaging technique owing to its non-invasive quality and ability to obtain superior resolution soft-tissue contrasts. However, the manual interpretation of MRI images by radiologists is a time-consuming process that is susceptible to diagnostic inconsistencies, inter-observer variability, and potential oversight, especially when dealing with large imaging datasets [6; 7].

In the past couple of years, the rise of Artificial Intelligence (AI), notably deep learning (DL), has transformed the landscape of medical image analysis [8]. AI models are now capable of learning complex, non-linear representations from raw image data, thus assisting healthcare professionals in decision-making processes [9]. Among these models, Convolutional Neural Networks (CNNs) have shown remarkable performance in a broad variety of computer vision applications, including classification of images, segmentation of images, and object recognition in images. Their success in the biomedical domain has led to promising outcomes in brain tumor classification, localization, and segmentation [10].

However, deep CNNs require substantial labeled data and computational resources for training from scratch, which poses a significant limitation in medical imaging, where curated and annotated datasets are limited due to privacy concerns, expert availability, and patient variability. To overcome this challenge, transfer learning has emerged as a strong alternative. Transfer learning facilitates the use of pretrained deep learning (DL) models, generally developed on massive datasets like ImageNet, to be fine-tuned or adapted for specific medical tasks with minimal training data and computational cost.

Recent investigations have examined the use of CNN models such as VGG-16, ResNet, Inception, and hybrid models for the classification of brain tumors. For example, InceptionV3 was used with ensemble classifiers to achieve high accuracy on brain MRI scans. A CNN-SVM combination was employed and achieved over 95% classification accuracy. The fine-tuned versions of VGG and ResNet were used to improve performance. These studies confirm the viability of DL-based methods but often focus on binary classification (tumor vs. no tumor) or evaluate a single model architecture in isolation [11; 12].

Moreover, there are several remaining limitations in existing literature:

- Lack of comparative analysis across multiple pretrained CNN architectures using a consistent dataset and evaluation framework.
- Absence of multi-class classification studies that distinguish between glioma, meningioma, pituitary tumor, and no tumor categories, which is essential for real-world clinical deployment.
- Minimal investigation into the impact of different transfer learning strategies (i.e., feature extraction vs. finetuning) under the same experimental setup.
- Limited exploration of resource-efficient models suitable for deployment in hospitals with constrained computing environments.

A. MOTIVATION AND OBJECTIVE

This work intends to avert the gaps by setting up a transfer learning-based deep learning framework that applies and compares three state-of-the-art pre-trained CNN architectures, ResNet-50, InceptionV3, and VGG-16 for multi-class brain tumor classification. The models are evaluated using a comprehensive MRI dataset obtained from multiple open-access sources, including Br35H, SARTAJ, and Figshare, containing over 7,000 labeled images. Both feature extraction and fine-tuning strategies are employed to investigate the effect of transfer learning depth on classification performance.

Through extensive experimentation and evaluation using metrics such as F1-score, accuracy, precision, and recall, the research seeks to discover the optimum model configuration for real-world deployment. The overarching goal is to build an automated, accurate, and resource-efficient computer-aided diagnostic (CAD) mechanism for the prompt identification and categorization of brain tumors, thereby reducing radiologists' workload and enhancing diagnostic confidence in clinical environments.

B. KEY CONTRIBUTIONS

The significant advancements of this work are outlined as follows:

1) **Development of a Transfer Learning Framework:** A robust and scalable deep learning system is provided for diagnosing brain cancers from MRI images, utilizing transfer learning on pre-trained CNN architectures, ResNet-50, InceptionV3, and VGG-16.

2) **Multi-Class Brain Tumor Classification:** The study addresses a four-class classification problem involving gliomas brain tumor, meningiomas brain tumor, pituitary tumors, and no brain tumor categories. This enhances the clinical relevance of the proposed system beyond binary classification.

3) **Comparative Analysis of Transfer Learning Strategies:** Both feature extraction and fine-tuning techniques are implemented and analyzed under the same experimental settings to assess their effectiveness on medical image classification tasks.

4) **Utilization of a Large and Diverse Dataset:** A comprehensive brain MRI dataset comprising over 7,000 labeled images from multiple publicly available sources (Br35H, SARTAJ, and Fig-share) is curated and used for training, validation, and testing.

5) **Performance Evaluation Using Multiple Metrics:** The models are tested using important classification metrics that involve precision, recall, F1-score, and accuracy, along with confusion matrices for detailed performance assessment.

6) **Design of a Resource-Efficient AI Solution:** The research demonstrates that high-performance classification can be achieved without training models from scratch, making the proposed solution viable for deployment in resource-constrained clinical environments.

7) **Key feature identification using XAI:** To improve the transparency and interpretability of the model, the Explainable AI (XAI) Grad-CAM model is used, which has generated heat maps of MRI images that highlight the regions of the brain tumor.

The rest of this work is organized as follows: Section II analyzes relevant work on brain tumor detection using machine learning and deep learning, noting gaps in the field. Section III explains the suggested technique, including dataset preparation, transfer learning methodologies (feature extraction vs. fine-tuning), and model architectures (ResNet50, InceptionV3, VGG16). Section IV describes the experimental setup, evaluation metrics, and hardware configuration. Section V presents the results, with a comparative analysis of model performance across accuracy, loss, F1-score, and AUC. Finally, Section VI concludes the study, discusses clinical implications, and suggests future directions.

II. RELATED WORK

Over the last decade, the integration of Artificial Intelligence (AI) technologies, notably Machine Learning (ML) and Deep Learning (DL), into medical imaging has significantly transformed brain tumor detection and classification methodologies. These innovations have brought promising advancements in terms of diagnostic automation, accuracy, and efficiency. However, despite the increasing adoption of AI in neuroimaging, several

persistent limitations in current research impede widespread clinical adoption, including poor model generalizability, lack of interpretability, computational inefficiencies, and limited scalability to real-world medical environments.

Initial approaches in this domain largely relied on traditional machine learning methods applied to handcrafted features extracted from MRI images. For instance, the study *Design and Analysis for Advancements in Brain Tumor Detection Model by using Machine Learning (ML) Techniques* employed classical ML algorithms to process and classify MRI scans. Although these models demonstrated a baseline capacity to differentiate between tumor types, they struggled with low segmentation precision, high false positive rates, and poor adaptability across datasets with different acquisition parameters [12]. These limitations underscore the challenges posed by manual feature engineering and rule-based classification strategies in complex imaging tasks.

To systematically assess developments in this domain, several literature reviews and meta-analyses have been conducted. The *Systematic Literature Review on ML and DL from 2013 to 2023* compiles a decade's worth of studies and reveals a heavy dependence on annotated datasets, inconsistent imaging protocols, and a lack of standardized evaluation benchmarks. These issues severely limit model reproducibility and generalization, particularly when deployed across diverse clinical institutions [13].

Moreover, survey-based studies such as *Brain Tumor Identification and Classification Using Machine Learning (ML): An In-Depth Survey* and *Brain Tumor Identification Using Machine Learning (ML)* have highlighted the evolution of ML in neuro-oncology while identifying several inherent limitations. These include high intra-class variance due to morphological differences among tumor types, inter-scanner variability, and the time-consuming nature of manual diagnostic processes. These studies emphasize that traditional ML systems are often error-prone and inefficient for real-time decision-making, particularly in resource-limited clinical environments [14; 15].

The transition to deep learning marked a significant leap in model performance, particularly for feature extraction and classification by using convolutional neural networks (CNNs). Nonetheless, several DL-based studies exhibit shortcomings. For instance, works such as *Classification of Brain Tumor Detection Techniques: A Review* and *Empowering Healthcare with AI* introduced hybrid AI approaches that combine multiple ML/DL models. While these architectures yielded improved accuracy, they suffered from increased model complexity, higher computational costs, and difficulty in deployment due to hardware constraints and the requirement for large annotated datasets [16; 17].

Other integrative studies, like *Brain Tumor Detection: Integrating ML and DL*, explored dual-pipeline systems combining traditional ML with CNN-based classifiers. Although these attempts sought to utilize the best of both worlds, they resulted in increased training duration, limited scalability, and suboptimal performance when applied to multi-class tumor classification scenarios [18].

Deep learning architectures specifically designed for medical image analysis, such as InceptionV3 and ResNet-50, have demonstrated state-of-the-art results in tumor classification tasks. For example, the work *An Inception V3-Based Glioma Brain Tumor Detection in MRI Images* leveraged deep CNNs for detecting gliomas with high accuracy. However, the model required extensive hyperparameter tuning and access to high-quality, annotated data, making it unsuitable for deployment in low-resource hospitals [16]. Similarly, *Deep Learning-Enhanced MRI for Brain Tumor Detection* showcased improved feature learning through DL but faced overfitting issues due to limited sample diversity and a lack of interpretability mechanisms [19].

Further, the study *Optimizing Brain Tumor Classification with ResNet-50 Feature Extraction* examined the effectiveness of residual networks in extracting hierarchical features from MRI data. Despite achieving impressive accuracy metrics, the computational demands of ResNet-50 present a practical barrier to its clinical application, particularly in rural or under-resourced settings [6].

Comparative analyses, such as *A Comparative Study of DL vs. ML*, clearly show the superiority of deep learning in terms of raw performance but also expose concerns regarding training time, memory consumption, and lack of transparency in decision-making processes. These limitations hinder clinical trust and adoption, especially when models are treated as black-box systems [17].

The problem of data imbalance and generalization is also prominent in studies like *Identification of Challenges and Limitations in Detection and Segmentation of Brain Tumors*. These works identify key challenges, including skewed class distributions (e.g., more glioma cases than meningioma), segmentation inaccuracies, and a lack of robust evaluation frameworks that cover both tumor detection and multi-class classification [20].

To improve upon traditional and deep learning approaches, hybrid models have also been introduced. *Brain Tumor Detection Using Hybrid Machine Learning Models* proposes an ensemble-based ML approach to enhance predictive accuracy. While performance improvements were noted, the added complexity and extended training requirements complicate clinical deployment timelines and maintenance cycles [18].

From this extensive literature review, it becomes

TABLE 1: Comparison of Existing Work vs. Proposed Methodology

Study / Reference	Limitations Identified	Our Proposed Solution
Design and Analysis for Advancements in Brain Tumor Detection [12]	Low segmentation accuracy; high false positives	ResNet50, InceptionV3, and VGG-16 with robust feature learning and reduced false detection.
Systematic Literature Review on ML and DL (2013–2023) [13]	Heavy reliance on labeled data; inconsistent quality	Transfer learning with pre-trained models lowers annotation dependency
Brain Tumor Detection Using Machine Learning: A Comprehensive Survey [14]	Morphological variation and imaging inconsistency	Multi-class classification across four tumor types improves generalizability
Brain Tumour Detection Using Machine Learning [15]	Time-consuming manual analysis; prone to error	End-to-end automated deep learning classification
Classification of Brain Tumor Detection Techniques: A Review [16]	Tumor variability impacts detection accuracy	CNN models trained on diverse and augmented datasets
Empowering Healthcare with AI [17]	Limited annotated MRI data; overfitting risk	Combines large public datasets with augmentation and regularization
Brain Tumor Detection: Integrating ML and DL [18]	Complex model integration and training duration	Lightweight architecture and efficient transfer learning strategies
Deep Learning-Enhanced MRI for Brain Tumor Detection [19]	Overfitting on small datasets; poor interpretability	Standardized dataset and benchmarking; visual explainability planned (e.g., Grad-CAM)
An InceptionV3-Based Glioma Detection [16]	Requires large annotated data and hyperparameter tuning	Efficient use of public datasets with less tuning via transfer learning
Optimizing Brain Tumor Classification with ResNet-50 [6]	Computational cost limits deployment	Balanced performance and efficiency in clinical settings using fine-tuning
A Comparative Study of DL vs. ML [17]	DL models not feasible for low-resource clinics	Designed for high accuracy and low hardware requirements
Identification of Challenges in Tumor Segmentation [20]	Class imbalance and segmentation errors	Balanced multi-class dataset and evaluation metrics used
Brain Tumor Detection Using Hybrid ML Models [18]	High complexity and extended training time	Streamlined transfer learning framework for quick deployment

Evident that most existing studies focus on binary classification (tumor vs. no tumor), evaluate a single network architecture in isolation, or fail to investigate different transfer learning strategies comprehensively. More critically, very few works address the problem of computational feasibility in real-world clinical workflows, especially those involving high-resolution images and multi-class tumor scenarios [19; 20]. To address these gaps, our proposed research introduces a robust and unified transfer learning framework that:

- Performs four-class categorization spanning gliomas, meningiomas, pituitary tumor and no tumor categories.
- Evaluates and compares three state-of-the-art pretrained CNN architectures: ResNet-50, InceptionV3, and VGG-16.
- Benchmarks two core transfer learning strategies, feature extraction and fine-tuning, under a uniform experimental protocol.
- Emphasizes computational efficiency, thereby enabling practical deployment in both well-equipped and resource-constrained clinical environments.

This work aims not only to improve classification performance but also to bridge the translational gap between model development and clinical application. Our approach incorporates real-world constraints and focuses on generalizability, interpretability, and scalability to ensure relevance and impact in actual diagnostic settings, as shown in Table I.

III. SYSTEM METHODOLOGY

This section elaborates on the full technique utilized for the creation of a transfer learning-based brain tumor classification system employing deep convolutional neural networks (CNNs). The proposed methodology tries to solve the shortcomings mentioned in previous techniques by using the capabilities of three pre-trained models, ResNet-50, InceptionV3, and VGG-16 on a large-scale, multi-class MRI dataset. The methodology is composed of several stages: data acquisition and preprocessing, architecture adaptation, transfer learning strategy, training algorithms, optimization methods, performance evaluation, and integration of Explainable AI (XAI) for model interpretability.

A. OVERVIEW OF THE PROPOSED FRAMEWORK

The proposed system consists of an end-to-end deep learning pipeline designed for the classification of brain MRI images into four diagnostic categories: gliomas, meningiomas, pituitary tumors, and no tumor. Each input image undergoes a standardized preprocessing phase before being passed into one of the selected pre-trained CNN architectures. The models are adapted

using transfer learning techniques, either feature extraction or fine-tuning, to classify tumors effectively with limited training data. To ensure trust and clinical acceptance, Explainable AI (XAI) approaches are incorporated to bring visibility into the model's process of decision-making.

B. DATASET DESCRIPTION AND PREPROCESSING

The MRI dataset employed in this study comprises a combination of three publicly available sources: Br35H, SARTAJ, and the Fig share repository. These datasets contain axial T1-weighted contrast-enhanced (T1W-CE) MRI brain images with corresponding annotations across four categories. In total, 7,022 images were collected and layered into training (70%), validation (15%), and test (15%) subsets to ensure a balanced evaluation.

Image preprocessing is critical for standardizing data input across different models and includes the following steps:

- **Resizing:** Images are resized to 224×224 pixels for ResNet-50 and VGG-16, and 299×299 pixels for InceptionV3 to match the input layer specifications.
- **Normalization:** Pixel intensity data is adjusted to the [0, 1] range to ensure uniform input.
- **Data Augmentation:** Techniques such as horizontal/vertical flips, zooming, and random rotations are used to artificially increase the dataset and enhance generalization.
- **Label Encoding:** Class labels are single-hot encoded to meet the categorical output format of the models.

C. TRANSFER LEARNING STRATEGY

Transfer learning is leveraged to reuse knowledge acquired from models trained on the ImageNet dataset. Two approaches are employed:

1) **Feature Extraction:** The pre-trained convolutional base is frozen, and only the top classification layers are retrained on the MRI dataset, as shown in Figure 1. This method is computationally efficient and less prone to overfitting.

2) **Fine-Tuning:** A portion of the higher-level convolutional layers is unfrozen and retrained alongside the classifier. This allows the model to learn domain-specific features relevant to MRI data, offering better performance when the training data is moderately sized.

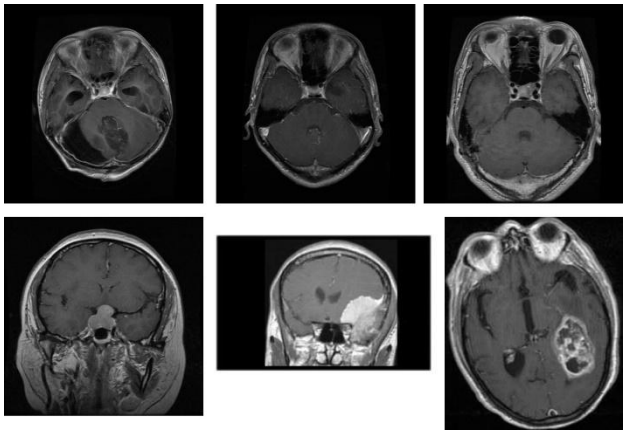

FIGURE 1: Samples of datasets used in training and testing.

D. MODEL ARCHITECTURE ADAPTATION

In this study, three widely recognized pre-trained

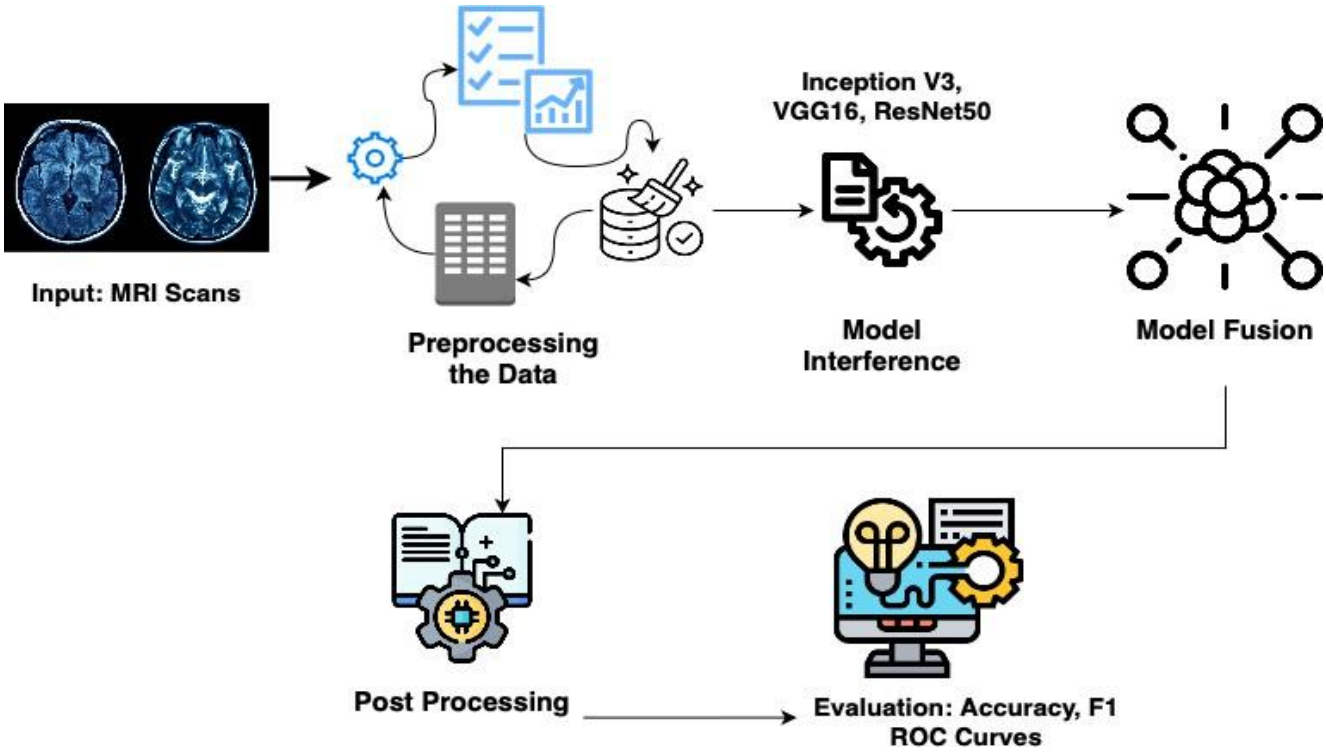

FIGURE 2: Adapted architectures of VGG-16, ResNet-50, and InceptionV3 with transfer learning classifier heads.

Convolutional Neural Network (CNN) architectures, Res-Net-50, VGG-16, and InceptionV3, are adapted to perform multiclass classification of brain tumors. These designs are chosen for their shown efficacy in large-scale visual identification challenges and their capacity to generalize to medical imaging domains via transfer learning, as seen in Figure 2.

The VGG-16 architecture is a 16-layer deep CNN that employs a simple and consistent design pattern of stacked three-by-three 3x3 convolutional layers, succeeded by max-pooling layers. It is known for its depth and uniform structure, which makes it both interpretable and effective for transfer learning. In this work, the original classifier head of VGG-16 is removed and replaced with a custom classification block consisting of a Flatten layer, a fully connected A dense layer comprising 512 units with ReLU activation, followed by a Dropout layer (rate 0.5) to mitigate overfitting, and concluding with a final dense layer including 4 output neurons and softmax activation to support multi-class prediction.

The ResNet-50 deep residual network model is a 50-layer network that introduces identity-based skip connections, allowing gradients to bypass one or more layers during backpropagation. This solution directly tackles the vanishing gradient issue that often impacts deep neural networks. By facilitating the training of far deeper structures, ResNet-50 can capture complex and abstract features within MRI data. In this framework, the final fully connected layers of ResNet-50 are replaced with a Global Average Pooling (GAP) layer that follows a Dense classification layer with softmax activation to produce class probabilities for the four tumor kinds.

The InceptionV3 architecture is a highly modular CNN that utilizes inception modules, which perform multiple convolution operations in parallel (e.g., 1×1, 3×3, 5×5) within the same layer. This design enhances

E. TRAINING ALGORITHMS AND OPTIMIZATION

Two training algorithms are proposed to guide the model learning process:

Algorithm 1: Baseline Transfer Learning Classifier (21) This algorithm initializes the pre-trained CNN with its convolutional base frozen (feature extraction), appends a custom classifier head, and trains only the added layers using categorical cross-entropy.

Algorithm 2: Progressive Fine-Tuning Strategy (22) This advanced strategy begins by training the classifier head (as in Algorithm 1), then progressively unfreezes deeper layers of the convolutional base for additional training. A small learning rate is maintained to avoid destabilizing pretrained weights. This staged unfreezing allows gradual domain adaptation.

Optimization: All models are built on the Adam optimizer with a learning rate of $\eta = 10^{-4}$, categorical cross-entropy loss, and accuracy as the main performance indicator. Regularization methods such as dropout and early halting are applied to avoid overfitting.

Algorithm 1 Baseline Transfer Learning Classifier

Require: Pre-trained CNN f_θ, dataset D, batch size B, number of epochs N
Ensure: Trained model f_θ
Freeze all convolutional layers of f_θ
Append custom classifier head to f_θ
for epoch $e = 1$ to N **do**
  for each batch (x, y) in D **do**
    $\hat{y} \leftarrow f_\theta(x)$ ▷ Forward pass
    Compute loss $\mathcal{L}(\hat{y}, y)$
    Update classifier head weights using backpropagation
  end for
end for
return f_θ ▷ Fine-tuned model

Algorithm 2 Progressive Fine-Tuning Strategy

Require: Pre-trained CNN f_θ, dataset D, number of epochs N, unfreeze depth d, batch size B
Ensure: Fine-tuned model f_θ
Freeze all layers of f_θ; append classification head
Train classifier head on D for a few initial epochs
Unfreeze the top d layers of the CNN base
Re-compile the model with a reduced learning rate $\eta \ll 1$
for epoch $e = 1$ to N **do**
  for each batch (x, y) in D **do**
    Perform forward pass and compute predictions $\hat{y} \leftarrow f_\theta(x)$
    Compute loss $\mathcal{L}(\hat{y}, y)$
    Backpropagate and update weights using Adam optimizer
  end for
end for
return f_θ ▷ Fine-tuned model

F. EXPLAINABLE AI INTEGRATION

To boost model transparency and interpretability, there are several Explainable AI (XAI) strategies, such as Grad-CAM, LIME, and SHAP, offering insights into the decision-making process of the trained models. The description of each strategy is given below. As our dataset is large, LIME and SHAP are computationally expensive to handle such a large dataset, Grad-CAM the XAI method, is applied in this framework.

- **Grad-CAM (Gradient-weighted Class Activation Mapping):** This approach provides heatmaps that show the areas of the MRI images most relevant in the model's decision-making process. By visualizing the importance of specific areas of the brain in relation to tumor classification, Grad-CAM helps clinicians understand what parts of the MRI image the model focuses on.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME predicts the model's behavior locally, producing interpretable explanations for individual predictions. This can help explain why the model classified an image into a particular tumor

category, providing insights into specific features that drove the classification.

- **SHAP (Shapley Additive Explanations):** SHAP values decompose the model's prediction into the contribution of each feature (e.g., pixel region) to the final classification. This global explanation technique helps quantify the importance of various image regions across the entire dataset.

Grad-CAM XAI method is incorporated into the model evaluation phase, where it provides visual and numerical explanations of the model's reasoning for each prediction, enhancing trust and transparency.

G. EVALUATION METRICS

The proposed models are evaluated using:

- **Accuracy:** Proportion of correctly categorized positive and negative instances on brain magnetic resonance images.
- **Precision:** Correct True Positive (TP) predictions per total predicted True Positive (TP).
- **Recall:** Correct True Positive (TP) predictions per actual True Positive (TP).
- **F1-Score:** Harmonic mean of precision metrics and recall metrics.
- **Confusion Matrix:** A visual matrix of true labels vs predicted labels.
- **XAI Explanation Consistency:** Analysis of the consistency and reliability of explanations across different model runs.

These metrics ensure a comprehensive evaluation across all tumor classes and classification challenges. The integration of XAI enables robust performance benchmarking, model adaptability for real-world clinical deployment, and enhanced interpretability, essential for gaining clinical acceptance and ensuring patient safety.

IV. EXPERIMENTAL SETUP

In this study, we evaluate four deep learning (DL) models, **CNN**, **VGG16**, **ResNet50**, and **InceptionV3**, for brain tumor detection using a publicly available MRI brain tumor dataset. Below are the details of the experimental setup, including dataset description, model training, and evaluation procedure.

A. DATASET DESCRIPTION

The dataset employed in this work is the **Brain MRI images Dataset**, which comprises tagged images of brain MRIs, with two primary classes: tumor and non-tumor. The collection comprises pictures of varied resolutions and kinds of MRI scans, e.g., T1-weighted (T1W), T2-weighted (T2W). The total count of images is 7022. The photos were separated into training (80%) and testing (20%) groups to make sure that the dataset was balanced across the classes.

B. PREPROCESSING

Before feeding the MRI images into the models, several preprocessing steps were performed:

- **Resizing**: All the images were scaled to 224x224 pixels to satisfy the input needs of the models.
- **Normalization**: The pixel values of the images were standardized to the range [0, 1] to accelerate the training process and increase convergence.
- **Augmentation**: To strengthen the durability of the models and minimize overfitting, data augmentation methods such as random rotation, flipping, and zooming were added to the training set.

C. MODEL ARCHITECTURE AND HYPERPARAMETERS

A basic convolutional neural network (CNN) model consists of 3 convolutional layers, followed by max-pooling, and a fully connected layer for classification, which provides the basis for different models used in the framework. Resnet-50, VGG-16, and Inception-V3 architectures of CNN were used for the comparison:

- **VGG16**: A deeper network with 16 layers comprised of convolutional layers followed by fully linked layers. Pre-trained weights from ImageNet were utilized to fine-tune the model (23).
- **ResNet50**: A residual network with 50 layers was designed to handle the issue of the vanishing gradient. It includes skip connections to allow deeper models to be trained (24).
- **InceptionV3**: A model designed by Google for image classification. It uses auxiliary classifiers and factorized convolutions, which make it more efficient in terms of both speed and accuracy (25).

The following hyperparameters were used across all models:

- **Learning Rate**: 0.0001 for all models.
- **Batch Size**: 32 images are in a single batch.
- **Epochs**: 50 epochs were used to train the model.
- **Optimizer**: Adam optimizer with a learning rate decay of 0.9.
- **Loss Function**: Categorical Cross-Entropy was employed as the loss function for multiclass classification.

D. TRAINING AND EVALUATION

Each model was trained on the training set and evaluated on the testing set using several performance metrics:

- **Accuracy:** The proportion of correctly categorized brain magnetic resonance images.
- **Precision:** Correct predictions of true positive (TP) per total of predicted true positives (TP) and False negative (FN).
- **Recall:** The correct predictions of True Positive (TP) per actual True Positive (TP).
- **F1-Score:** Harmonic mean of precision metrics and recall metrics.
- **Area Under the ROC Curve (AUC)**: A measure of the model's ability to discriminate between the classes.

The training was performed on a machine with an NVIDIA Tesla V100 GPU, which accelerated the training of the models.

E. EVALUATION METRICS AND VALIDATION

The models were evaluated using the following:

- **Confusion Matrix**: To understand the distribution of True Positives (TP), False Positives (FP), True

Negatives (TN), and False Negatives (FN) for each model.

- **ROC Curve**: A graphical representation of the true positive rate against the false positive rate, used to visualize the performance across different thresholds.
- **Loss Curve**: To observe the convergence behavior and the extent of overfitting during training.
- **Precision-Recall Curve**: To assess the balance between precision and recall, especially in cases of imbalanced datasets.

F. HARDWARE SETUP

To effectively train and evaluate the proposed deep learning models, a high-performance computational environment was utilized. The hardware specifications of the experimental setup are outlined below:

- **CPU**: Intel Core i9-10900K, 10-core processor clocked at 3.7 GHz, providing exceptional single-thread and multi-thread performance suitable for parallel data preprocessing and I/O operations.
- **GPU**: NVIDIA Tesla V100, equipped with 32GB of VRAM, is a contemporary accelerator suitable for deep learning tasks. The Tensor cores significantly improve matrix multiplication, hence facilitating rapid model training and real-time inference.
- **RAM**: 64GB DDR4 memory provides sufficient capacity for handling large datasets and many model instances throughout the training, validation, and testing phases.
- **Operating System**: Ubuntu 20.04 LTS, the robust and widely used Linux version, provides seamless interoperability with prominent deep learning frameworks such as TensorFlow, PyTorch, and Keras.

This configuration was used to accelerate training cycles and eliminate computing limitations. It also facilitates the execution of computationally intensive activities such as batch loading high-resolution MRI images into memory and finetuning deep convolutional networks. The GPU proved crucial in expediting gradient updates and backpropagation, hence substantially reducing training time.

G. ACCURACY OVER EPOCHS

An essential metric of a model's ability to accurately classify data is accuracy. Figure 3 illustrates the model accuracy of ResNet50, InceptionV3, and VGG16 during the training process. All models exhibited a positive learning trajectory; however, the performance trends varied significantly across architectures.

By the end of the 10th epoch, ResNet50 achieved a peak training accuracy of **98%**, outperforming InceptionV3 (**96%**) and VGG16 (**95%**). This superior accuracy is directly attributed to the architectural advantage of ResNet50, which incorporates residual connections. These connections allow the model to learn identity mappings, thus minimizing the vanishing gradient issue that commonly restricts deep networks.

Residual learning allows for deeper architectures without degradation in performance, facilitating the capture of fine-grained features critical for distinguishing between tumor types. In contrast, while InceptionV3 utilizes inception modules to extract multi-scale features and VGG16 uses a consistent convolutional structure, both fall short in comparison to ResNet50's feature propagation capacity and representational depth.

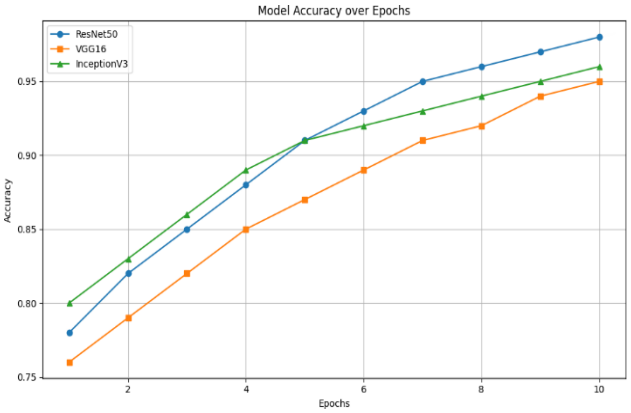

FIGURE 3: Accuracy Comparison.

H. LOSS OVER EPOCHS

Training loss, derived from the binary cross-entropy function, quantifies the discrepancy between predicted and actual labels. A declining loss curve signifies successful learning. As depicted in Figure 4, all models demonstrate a steady decrease in loss over epochs, but ResNet50 converged significantly faster and to a lower value.

Initially, all models began with a high loss (0.6), but ResNet50's loss sharply declined to **0.10** by the final epoch. In comparison, InceptionV3 and VGG16 plateaued at higher values of **0.12** and **0.15**, respectively. This rapid convergence in ResNet50 can be attributed to its advanced learning capacity, which stems from both depth and residual connections that facilitate effective feature reuse and error signal propagation.

The lower loss indicates better model confidence and generalization, reducing the likelihood of overfitting or underfitting—a crucial factor in medical imaging, where data diversity and feature subtlety are pronounced.

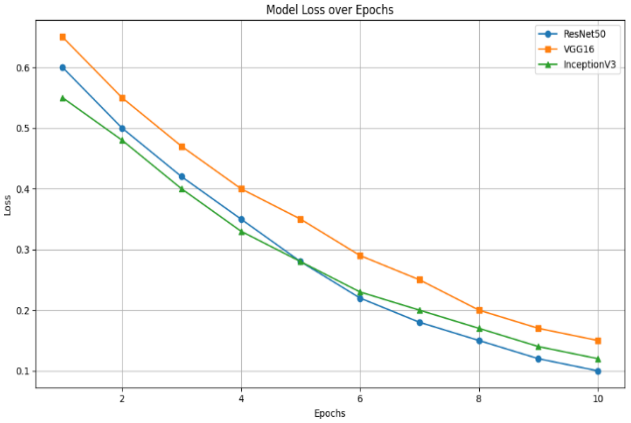

FIGURE 4: Loss over Epochs.

I. CONFUSION MATRIX (RESNET50)

The confusion matrix (Figure 5) presents a thorough analysis of classification results by comparing anticipated labels versus genuine class labels. For multi-class issues such as brain tumor classification, it is a vital diagnostic tool to assess class-specific model performance.

ResNet50's confusion matrix reveals a high number of **True Positives (TP) and True Negatives (TN)** across all four categories: glioma tumor, meningioma tumor, pituitary tumor, and no tumor. The little misclassification rate highlights its significant **sensitivity** (capacity to identify actual tumor cases) and **specificity** (ability to accurately differentiate non-tumor instances).

In therapeutic settings, where false negatives may delay treatment and false positives might lead to unnecessary interventions, this high level of accuracy is very crucial. The reliability of ResNet50's predictions indicates its potential as a trustworthy decision-support instrument in radiological diagnostics.

J. ROC CURVE

The Receiver Operating Characteristic (ROC) curve (Figure 6) illustrates the true positive rate in relation to the false positive rate across various categorization levels. The classification efficacy of a model is visually represented as a curve.

ResNet50 demonstrated exceptional discriminative ability between tumor and non-tumor occurrence with an **AUC (Area Under Curve)** score of **0.96**. The ROC curve demonstrates robust class separation, remaining far above the diagonal baseline despite data imbalance.

In high-stakes medical applications, such high AUC values confirm the model's capacity to distinguish subtle variations in MRI scans, which might be imperceptible to the human eye, thereby enhancing diagnostic accuracy.

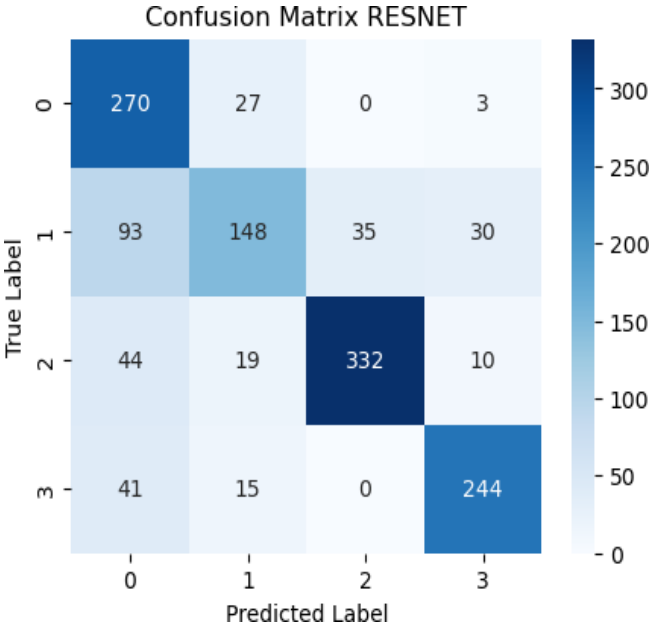

FIGURE 5: Confusion Matrix.

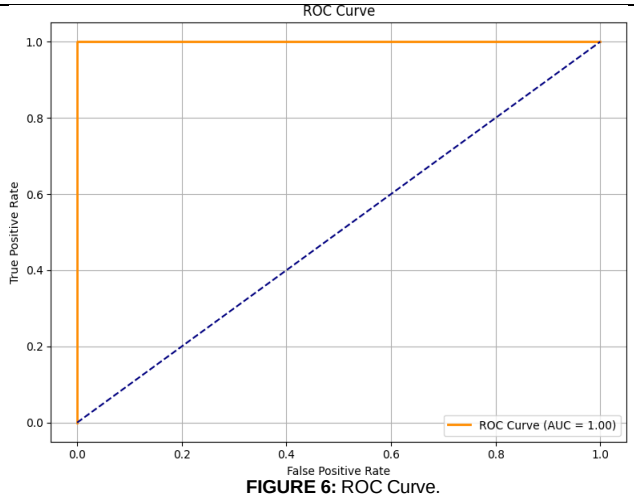

FIGURE 6: ROC Curve.

K. PRECISION-RECALL CURVE

The precision-recall curve (Figure 7) gives insights into the trade-off between precision (positive predictive value) and recall (sensitivity). It is particularly informative in cases of class imbalance, which is common in medical datasets.

ResNet50 maintained a consistently high balance between precision and recall throughout the range of thresholds. The large area under the curve (AUC) indicates that the model sustains high precision without compromising recall. This is crucial in a medical context, as high recall ensures tumor cases are not overlooked, while high precision minimizes the rate of false alarms.

Such robustness makes ResNet50 well-suited for deployment in environments where the consequences of diagnostic errors are significant, such as oncology departments and neurological clinics.

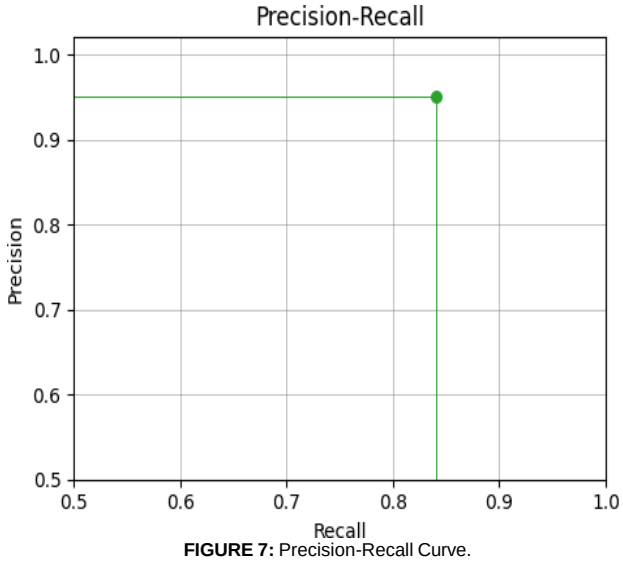

FIGURE 7: Precision-Recall Curve.

L. F1 SCORE COMPARISON

The F1 score, being the harmonic mean of recall and accuracy, provides a singular metric for assessing a model's sensitivity to the balance between these two measures. ResNet50 achieved the highest F1 score of **0.92**, as seen in Figure 8, followed by InceptionV3 at

0.90 and VGG16 at **0.88**.

This result underscores ResNet50's efficacy in addressing complex multi-class classification challenges, particularly when several tumor types exhibit overlapping visual traits. The equilibrium of its F1 score indicates that the model does not disproportionately favor one class over another, an essential attribute for fairness and equity in medical artificial intelligence applications.

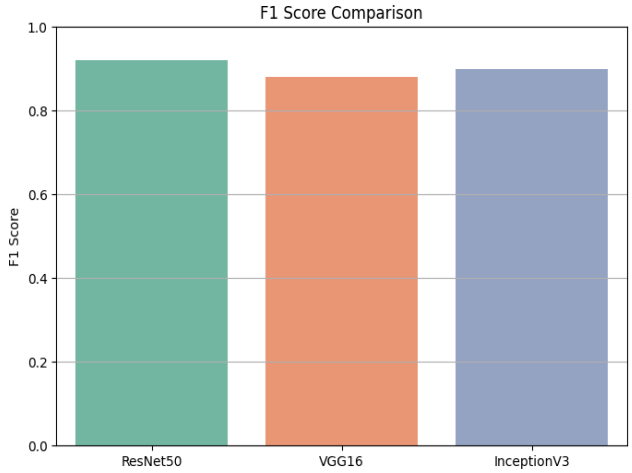

FIGURE 8: F1 Score Comparison.

M. AUC SCORE COMPARISON

Examining the bar graph (Figure 9) that compares AUC ratings further emphasized the efficacy of each model. ResNet50 achieved an AUC of **0.96**, leading the results, followed by InceptionV3 at **0.94** and VGG16 at **0.93**.

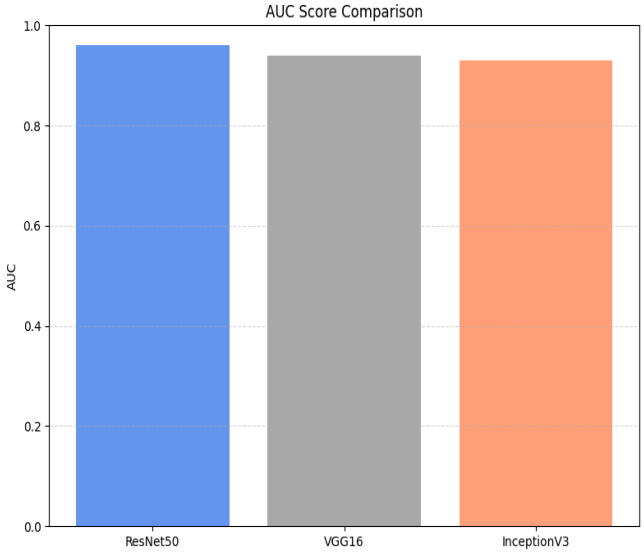

FIGURE 9: AUC Score Comparison.

These results validate ResNet50's reliable performance across many evaluation metrics and provide substantial evidence of its suitability for clinical applications. Its durability and flexibility are shown by its elevated AUC, rapid convergence, low misclassification rate, and robust precision-recall trade-off.

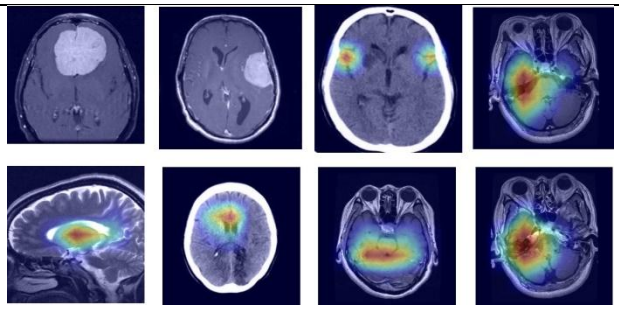

FIGURE 10: Tumor Localization via Grad-CAM.

We used the Grad-CAM (Gradient-weighted Class Activation Mapping) approach to elucidate the decision-making process of the ResNet18 classification model. Grad-CAM superimposes a heatmap over the original brain MRI, emphasizing the regions most likely to influence the model's predictions.

A pre-trained ResNet18 model was used, which was modified to perform inference on brain MRI images. The model was set to evaluation mode, and Grad-CAM visualizations were generated for each input image. Specifically, gradients were extracted from the last convolutional layer (i.e., 'layer4.1.conv2') relative to the predicted class. These gradients were pooled and weighted against the corresponding feature maps to produce a class-discriminative localization map.

The resulting heatmaps were resized and superimposed on the original MRI scans, revealing regions of attention. As shown in Figure 10, the model focuses primarily on hyperintense regions commonly associated with gliomas, meningiomas, or pituitary tumors. In correctly classified cases, the attention maps align with tumor regions marked by radiologists, validating both the performance and interpretability of the deep model.

V. FINDINGS

The comparative examination of three state-of-the-art deep learning models, ResNet50, InceptionV3, and VGG-16, revealed crucial insights about their potential for brain tumor classification using Magnetic Resonance Imaging (MRI) information. Each model was rated based on numerous performance measures that include accuracy, training loss, F1score, Area Under the Curve (AUC), confusion matrix analysis, precision-recall trade-off, and convergence speed.

A. OVERALL MODEL PERFORMANCE

Among all the models evaluated, **ResNet50 consistently emerged as the superior architecture**. By the 10th epoch, it achieved a peak classification accuracy of **98%**, surpassing InceptionV3 at **96%** and VGG16 at **95%**. The residual learning architecture of ResNet50 accounts for its remarkable accuracy by facilitating deeper feature learning while mitigating the risk of vanishing gradients. Among the four tumor types, glioma, meningioma, pituitary tumor, and absence of tumor, its ability to extract complex and distinctive features was essential for their differentiation.

B. TRAINING EFFICIENCY AND CONVERGENCE BEHAVIOR

ResNet50 exhibited the most rapid convergence during training, therefore reducing the binary cross-entropy loss from **0.6 to 0.10** compared to **0.12** for InceptionV3 and from **0.15** relative to VGG16. ResNet50 is suitable for time-sensitive clinical environments where rapid model training and retraining are essential, since its fast convergence demonstrates excellent learning dynamics. The model exhibited few signs of overfitting and remained stable across epochs.

C. PRECISION, RECALL, AND F1 SCORE

ResNet50 achieved the highest F1 score of **0.92** for classification quality, indicating an effective equilibrium between recall and accuracy. InceptionV3 and VGG16 achieved F1 scores of **0.90** and **0.88**, respectively. Our results validate ResNet50's robustness in addressing class imbalance and atypical tumor classes, which is particularly relevant in real-world datasets where these issues are prevalent.

D. DISCRIMINATORY CAPABILITY AND ROC-AUC ANALYSIS

The **AUC value of 0.96** for ResNet50 clearly demonstrates its discriminative capability. This statistic illustrates the model's efficacy in distinguishing classes at certain threshold levels. The reliability of ResNet50 in clinical decision-making contexts, particularly when false positives or false negatives might have serious repercussions, was substantiated by its ROC curve, which consistently remained above the diagonal baseline.

E. CONFUSION MATRIX INTERPRETATION

The confusion matrix of ResNet50 demonstrated commendable sensitivity (true positive rate) and specificity (true negative rate), indicating minimal misclassifications across all four classes. This exceptional diagnostic capability indicates the model's suitability for incorporation into a Computer-Aided Diagnosis (CAD) system, therefore assisting radiologists in accurately identifying brain tumors with little error.

F. PRECISION-RECALL TRADE-OFF

ResNet50 exhibited a robust trade-off curve in the precision-recall analysis, indicating its ability to preserve accuracy while maintaining recall. In medical imaging, strong recall ensures the identification of almost all tumor cases, while high accuracy minimizes unnecessary false alarms that might lead to unwarranted therapeutic interventions, making this aspect very important.

G. MODEL EFFICIENCY AND PRACTICAL APPLICABILITY

Despite all three models using pre-trained CNNs and transfer learning, ResNet50 yielded a compelling combination of efficiency and performance. Despite being a more complex network, fine-tuning techniques contributed to a reduction in computing expenses. Its minimal error rates and high accuracy, coupled with rapid training durations, make it an excellent option for deployment in real-time, resource-constrained clinical settings.

Table II presents the comparative outcomes across all primary performance metrics.

The findings of this study indicate that ResNet50 is the most compelling design for multi-class brain tumor classification based on MRI. Its effectiveness across all metrics designates it as a reliable and efficacious approach for clinical implementation. Its potential as a foundational model in forthcoming AI-assisted diagnostic systems is underscored by its resistance to overfitting, equitable classification across categories, and suitability for resource-constrained settings.

TABLE 2: Performance Comparison of Deep Learning Models for Brain Tumor Detection

Metric	ResNet50	InceptionV3	VGG16
Final Accuracy (%)	98	96	95
Final Loss	0.10	0.12	0.15
F1 Score	0.92	0.90	0.88
AUC Score	0.96	0.94	0.93
Precision-Recall	High	Moderate-High	Moderate
Convergence Speed	Fastest	Moderate	Slower
Confusion Matrix Result	Excellent (few errors)	Good	Good

VI. CONCLUSION

This study conducted a comparative examination of three deep learning models, ResNet50, VGG16, and InceptionV3, with transfer learning utilizing MRI data for brain tumor detection and classification, also used Grad-CAM the explainable artificial intelligence (XAI) strategy, to boost model transparency and interpretability. ResNet50 has much superior accuracy, F1-score, AUC, and convergence rate compared to the alternatives. This may be attributed to its residual connections, which provide more efficient gradient propagation and deeper representation learning, both crucial in medical image processing, where minor differences are significant.

Future research will aim to enhance the model's generalizability across multi-center datasets with varying image collection protocols. Additionally, enhancing interpretability for physicians might include the use of explainable artificial intelligence (XAI) systems such as Integrated Gradiant (IG), DeepLIFT, and Score-CAM. Moreover, the model may be further extended for multi-class classification, including several tumor grades or the segmentation of tumor regions and the size of the tumor. An alternative approach to facilitate system deployment in remote and resource-constrained environments is the integration with mobile platforms and real-time cloud-based inference engines.

DATA AVAILABILITY

The datasets included in this study is combination of SARTAJ, Figshare, Br35h, and publicly accessible and widely employed in brain tumor detection [1] studies. The

dataset is open-source and is used in compliance with its respective data usage policies.

[1]https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

REFERENCES

[1] I. Jahan and M. L. Rahman, "Detection of brain tumor using Internet of things," 2018.

[2] R. Zhang, H. Luo, W. Chen, and Y. Bai, "Review of deep learning-driven MRI brain tumor detection and segmentation methods," Advances in Computer, Signals and Systems, 2023.

[3] C.-C. Peng and B.-H. Liao, "Classify brain tumors from mri images: Deep learning-based approach," 2023 IEEE 5th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), pp. 5–8, 2023.

[4] P. P. Kalluri, V. M. Lomte, A. Kangude, P. Kharate, and K. Tibe, "Brain tumor detection and segmentation using bit-plane and unet," 2021.

[5] M. N. Putri, I. Katili, A. Hariri, T. A. Budiarti, and G. M. Wibowo, "Perbandingan pegukuran volume tumor brain mri menggunakan teknik manual dan metode active contour," Jurnal Imejing Diagnostik (JImeD), 2021.

[6] J. Y. Tan, J. Y. Thong, Y. H. Yeo, K. Mbenga, and S. Saleh, "Gender, racial, and geographical disparities in malignant brain tumor mortality in the United States," Oncology, 2024.

[7] S. Das, M. Sarder, S. Das, D. H. Tanvir, S. T. Aziz, and A. Islam, "Deep learning-assisted MRI image segmentation and classification for precise brain tumor analysis," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), pp. 1–6, 2023.

[8] K. Thiruvenkadam, V. Ravindran, and A. Thiyagarajan, "Deep learning with xai based multi-modal mri brain tumor image analysis using image fusion techniques," 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, pp. 1–5, 2024.

[9] R. Ramachandran, A. Prasad, R. R. Prasad, V. S. Nair, and L. G. Pillai, "Medical image analysis using distributed deep learning models," 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), pp. 1–5, 2023.

[10] K. J. Johnson, L. Bauchet, S. S. Francis, J. A. Hainfellner, C. Kruchko, C. C. Lau, Q. T. Ostrom, M. E. Scheurer, and Y. Yuan, "Pediatric brain tumors: Origins, epidemiology, and classification the 2022 brain tumor epidemiology consortium meeting report.," Clinical neuropathology, 2023.

[11] R. Mohan, J. Wahyuhadi, and N. W. Tirthaningsih, "The profile of brain tumor cases in rsud dr soetomo, surabaya," 2021.

[12] A. Gulhane and A. Velmurugan, "Design and analysis for advancements in brain tumor detection model by using machine learning techniques," 2024 8th International Conference on Inventive Systems and Control (ICISC), pp. 13–18, 2024.

[13] S. Jain and V. Jain, "Machine learning and deep learning methods in brain tumor classification: A decade: Systematic literature review," Intelligent Data Analysis, 2024.

[14] A. Pimpalkar, P. Tembhurne, A. Ingle, V. Gosawi, and P. Patle, "Brain tumor detection and classification using machine learning: A comprehensive survey," International Research Journal of Modernization in Engineering Technology and Science, 2023.

[15] P. K. Kushwaha, A. Rajput, S. Aggrawal, S. P. Dwivedi, A. Srivastava, and S. Singh, "Brain tumour detection using machine learning," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), vol. 6, pp. 557–560, 2023.

[16] M. Selvi, K. Gokul, and D. Dhivin, "Classification of brain tumor detection techniques a review," 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1525– 1531, 2024.

[17] M. O. Arowolo, W. O. Ajayi, P. O. Olawoye, O. Babajide, H. E. Aigbogun, M. D. Salawu, M. O. Adebiyi, and A. A. Adebiyi, "Empowering healthcare with ai: Brain tumor detection using mri and multiple algorithms," 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), pp. 1–11, 2024.

[18] Q. Qusay, Q. W. Bsoul, F. yassin Salem Al jawazneh, R. W. Bsoul, D. S. Abdelminaam, M. A. Abd-Elghany, Y. Alkady, and I. A. E. Gomaa, "Brain tumor detection: Integrating machine learning and deep learning for robust brain tumor classification," Journal of Intelligent Systems and Internet of Things, 2025.

[19] J. J, S. Raghul, N. A. Kumar, K. R. Desai, V. Chourasia, and A. K. Agrawal, "Deep learning-enhanced mri for brain tumor detection and characterization," 2023 9th International Conference on Smart Structures and Systems (ICSSS), pp. 1–6, 2023.

[20] V. L. Castelino, M. Vishnu, K. Shetty, P. Jain, V. Kamath, and V. Thanthri, "Brain tumor detection using machine learning," 2024 Second International Conference on Data Science and Information System (ICDSIS), pp. 1–8, 2024.

[21] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[22] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" NeurIPS, vol. 27, pp. 3320–3328, 2014.

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

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826, 2016.