Vol. 07, No. 01, pp. 110 –123 (2026)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

MRI-Based Brain Tumor Diagnosis Using Preprocessing Pipeline and Transfer Learning Models

Mahmood A. Nawar, Ahmed M. Dinar*

College of Computer Engineering, University of Technology- Iraq, mahmood.a.nawar@uotechnology.edu.iq, ahmed.m.dinar@uotechnology.edu.iq

**Correspondence: ahmed.m.dinar@uotechnology.edu.iq*

Abstract

Brain tumors represent a significant global health challenge, and their accurate and timely diagnosis is critical for effective treatment planning and improved patient outcomes. While transfer learning has shown promise in this domain, its performance is highly dependent on the quality of input data. This research introduces a novel, comprehensive 8-step preprocessing pipeline designed to significantly enhance the quality and feature visibility of Magnetic Resonance Imaging (MRI) scans for automated brain tumor classification. The pipeline includes resizing, grayscale conversion, Gaussian blurring, Otsu thresholding, contour isolation, Region of Interest (ROI) cropping, normalization, and a Power-law transform. To validate the efficacy of our proposed pipeline, we utilized a merged dataset of 10,287 MRI images from the Masoud and SARTAJ collections, encompassing four classes: glioma, meningioma, pituitary, and normal. This enhanced dataset was used to train and evaluate seven state-of-the-art transfer learning models: Xception, DenseNet-121, GoogLeNet, MobileNet, MobileNet-v2, VGG-19, and ResNet-50. Our rigorous preprocessing resulted in exceptional classification performance, with the Xception model achieving a peak accuracy of 98.68%. This study demonstrates that a meticulous and well-designed preprocessing pipeline is a critical and often overlooked component in developing highly accurate and reliable Computer-Aided Diagnosis (CAD) systems for clinical applications.

Keywords: *Brain tumor classification, Transfer learning, Deep learning, MRI, Image preprocessing, Computer-aided diagnosis, Convolutional neural networks.*

Received: December 12th, 2025 / Revised February 08th, 2026 / Accepted: February 15th, 2026 / Online: February 18th, 2026

I. INTRODUCTION

The human brain, a complex and highly specialized organ, serves as the central command center for the entire nervous system, orchestrating cognitive function, motor skills, and vital physiological processes. However, this delicate organ is susceptible to various pathologies, chief among them being brain tumors (BTs) [1]. Brain tumors are characterized by the abnormal and uncontrolled proliferation of cells within the brain tissue or its surrounding structures [2]. These growths can be broadly classified as benign (non-cancerous), which are typically slow-growing and localized, or malignant (cancerous), which are aggressive, rapidly growing, and often invasive [3].

The World Health Organization (WHO) has categorized brain tumors into over 100 distinct types, with common primary tumors including gliomas, meningiomas, and pituitary adenomas [3, 4]. Gliomas, arising from glial cells, are the most prevalent and vary widely in malignancy, with glioblastoma being the most aggressive form. Meningiomas originate from the protective membranes (meninges) surrounding the brain and spinal cord and are often benign but can cause complications due to pressure on the brain tissue. Pituitary tumors, mostly benign adenomas, can disrupt hormonal balance and affect vision [4]. The risks associated with brain tumors are profound, as the confined space within the skull means that even benign growths can exert pressure on critical brain structures, leading to a range of debilitating symptoms. These can include persistent headaches, nausea, seizures, sensory impairments, and progressive cognitive and motor function decline, severely impacting the patient's quality of life [4, 5].

The impact of a brain tumor on an individual is devastating, extending beyond the physical symptoms to significant psychological and socioeconomic burdens on patients and their families [4]. The aggressive nature of malignant tumors, coupled with the potential for permanent neurological damage, underscores the critical need for early and accurate diagnosis [6]. Globally, brain tumors represent a significant health challenge. Cancer is a leading cause of death worldwide, and brain tumors, while representing less than 2% of all cancers, are the tenth leading cause of mortality for both men and women [5, 7]. Annually, an estimated 126,000 new cases are diagnosed worldwide, with over 308,000 new instances reported in 2020

doi: 10.38094/jastt71744

alone [5, 7]. In the United States, it is projected that approximately 24,810 adults will be diagnosed with malignant brain tumors in 2024, resulting in an estimated 18,990 deaths [4].

The prognosis for malignant brain tumors is often grim, with the 5-year survival rate for adults remaining as low as 36% [3]. This highlights the urgency for improved diagnostic and therapeutic strategies. Early detection is paramount, as timely intervention before tumors advance greatly enhances treatment success and can significantly improve survival rates [4, 8].

Magnetic Resonance Imaging (MRI) stands as the gold standard for the non-invasive diagnosis and monitoring of brain tumors due to its exceptional capacity to generate detailed, high-resolution images of soft tissues [1], [9]. MRI provides crucial information regarding the tumor's location, size, shape, and internal structure, which is essential for pathology comprehension, treatment planning, and surgical guidance [1], [2]. However, the traditional diagnostic process, which relies on the manual interpretation of these complex MRI scans by radiologists, is inherently challenging. This manual process is time-consuming, laborious, and prone to inter- and intra-observer variability and human error, especially when dealing with the high volume and complexity of modern medical images [9], [10].

The structural complexity, high volatility, and significant variability in tumor size and shape across patients make manual segmentation and diagnosis a particularly tedious and difficult task [9]. The necessity for multiple hospital visits and extensive testing can also lead to delays in treatment initiation, which is detrimental for fast-growing malignant tumors [4]. Consequently, there is a crucial and urgent need for more reliable and trustworthy advanced detection techniques, commonly referred to as Computer-Aided Diagnosis (CAD) systems, to assist clinicians and improve the speed and accuracy of tumor classification [10].

The integration of Artificial Intelligence (AI) with Magnetic Resonance Imaging (MRI) has revolutionized the field of medical diagnostics, particularly for brain tumor classification. MRI is the gold standard for brain imaging due to its excellent soft-tissue contrast and non-invasive nature, providing detailed anatomical information without the use of ionizing radiation [11]. However, the manual interpretation of these images is time-consuming and prone to inter-observer variability. AI, specifically deep learning, has emerged as a powerful tool to overcome these limitations by enabling the automated analysis of MRI scans, leading to faster and more accurate diagnoses [12]. Deep learning models, particularly Convolutional Neural Networks (CNNs), can learn complex hierarchical features directly from image data, making them highly effective for tasks like tumor detection, segmentation, and classification.

Transfer learning, a key technique in deep learning, has further accelerated progress in this domain. Instead of training a deep neural network from scratch, which requires massive amounts of labeled data, transfer learning leverages pre-trained models that have been trained on large-scale image datasets like ImageNet. These models, such as VGG-16, ResNet-50, and Xception, have already learned a rich set of low-level features (e.g., edges, textures, shapes) that are transferable to medical imaging tasks. By fine-tuning these pre-trained models on brain tumor MRI datasets, researchers can achieve high classification accuracies even with limited data, significantly reducing training time and computational cost. This approach has consistently demonstrated state-of-the-art performance, with studies reporting accuracies in the range of 95-98% for multi-class brain tumor classification [11],[12].

The primary objective of this research is to introduce and validate a novel, multi-step image preprocessing pipeline designed to optimize MRI image quality for deep learning-based brain tumor classification. We hypothesize that our rigorous 8-step pipeline will significantly enhance feature visibility, leading to improved classification accuracy across a suite of state-of-the-art transfer learning models. To test this hypothesis, we will apply our pipeline to a large, merged dataset and conduct a comprehensive comparative analysis of seven pre-trained architectures to identify the most effective model for this task.

II. Related Works

The field of automated brain tumor classification from MRI scans has witnessed substantial progress through the application of deep learning methodologies. Contemporary research demonstrates a spectrum of approaches, ranging from lightweight architectures optimized for computational efficiency to sophisticated ensemble systems designed for maximum accuracy. A critical examination of recent literature reveals both the achievements and persistent limitations that motivate the present study.

The choice of backbone architecture represents a fundamental decision in brain tumor classification system design. Recent work has explored the trade-offs between model complexity and performance. Agrawal and Chaki [3] introduced CerebralNet, which leverages a MobileNetV2 backbone enhanced with Atrous Spatial Pyramid Pooling (ASPP) and Atrous Convolution blocks to capture multi-scale contextual information. Their approach achieved 96% accuracy on an augmented dataset, demonstrating that lightweight architectures can deliver competitive performance when augmented with sophisticated feature extraction mechanisms. This finding is particularly relevant for clinical deployment scenarios where computational resources may be constrained. However, the reliance on extensive probabilistic augmentation techniques raises questions about whether the model's performance stems primarily from architectural innovation or data augmentation strategies. In contrast, Maqsood et al. [9] adopted a hybrid strategy, combining a modified MobileNetV2 architecture for feature extraction with an entropy-based feature selection method and a multiclass Support Vector Machine (M-SVM) for final classification. This multi-modal approach achieved accuracies of 97.47% on the BraTS 2018 dataset and 98.92% on the Figshare dataset, suggesting that the integration of traditional machine learning classifiers with deep feature extractors can yield superior results compared to end-to-end deep learning alone. The authors also incorporated a custom 17-layer deep neural network for tumor segmentation, highlighting the importance of preprocessing and region-of-interest isolation in the classification pipeline.

Transfer learning from pre-trained models has emerged as a dominant paradigm, yet the selection of appropriate architectures and fine-tuning strategies remains an active area of investigation. Classical architectures such as VGG-19 have been extensively studied. Sajjad et al. [13] employed a fine-tuned VGG-19 model with cascade CNN architecture for segmentation, achieving 94.58% accuracy. Similarly, Swati et al. [14] reported 94.82% accuracy using VGG-19 on contrast-enhanced MRI images. More recently, Narayankar and Baligar [15] applied VGG-19 to a pooled dataset comprising Figshare, Br35H, and SARTAJ sources, achieving 95.11% accuracy. While these studies confirm the utility of VGG architectures, their performance plateaus below 96%, suggesting that the relatively shallow depth and simple convolutional structure of VGG-19 may limit its capacity to capture the complex hierarchical features present in brain tumor MRI scans. In contrast, deeper residual architectures have demonstrated superior performance. Kumar et al. [16] implemented ResNet-50 with global average pooling, achieving 97.48% accuracy. The residual connections in ResNet architectures facilitate gradient flow during training, enabling the learning of more discriminative features. Togacar et al. [17] proposed BrainMRNet, a custom architecture incorporating attention modules, which achieved 96.05% accuracy. While attention mechanisms can enhance feature selectivity, the modest performance gain suggests that architectural novelty alone may not be sufficient without corresponding advances in data preprocessing and augmentation strategies.

Ensemble and hybrid methodologies represent an alternative approach to maximizing classification performance by leveraging the complementary strengths of multiple models. Kibriya et al. [18] developed a feature fusion framework that extracts deep features from both GoogLeNet and ResNet-18, subsequently classified using SVM and KNN classifiers, achieving 97.7% accuracy. This approach capitalizes on the diverse feature representations learned by different architectures.

Haque et al. [4] advanced this concept further by proposing a stacking ensemble that combines EfficientNetB0, MobileNetV2, GoogLeNet, and a Multi-level CapsuleNet, using CatBoost as a meta-learner. Their system achieved F1-scores of 97.81% and 98.32% on two merged datasets (M1 and M2), demonstrating the power of sophisticated ensemble strategies. The authors also addressed class imbalance through Borderline-SMOTE and employed PCA combined with Gray Wolf Optimization for feature selection, illustrating the importance of comprehensive data preprocessing pipelines. However, ensemble approaches introduce significant computational overhead during both training and inference, and the increased model complexity may hinder interpretability and clinical adoption. Furthermore, the marginal performance gains achieved by these complex ensembles compared to well-tuned single models raise questions about the practical cost-benefit trade-off in real-world deployment scenarios.

A growing trend in the literature is the integration of Explainable AI (XAI) techniques to enhance the trustworthiness and interpretability of deep learning models, a critical factor for clinical adoption. For instance, Narayankar and Baligar [15] utilized Layer-wise Relevance Propagation (LRP) to provide pixel-wise relevance maps for their VGG-19 model, offering insights into the model's decision-making process.

Similarly, Agrawal and Chaki [3] incorporated LIME (Local Interpretable Model-agnostic Explanations) with their CerebralNet to explain its predictions. Other studies, such as Haque et al. [4], have also emphasized the importance of explainability in their ensemble frameworks. These works underscore a clear demand in the field: it is no longer sufficient for a model to be accurate; it must also be transparent. While these studies introduce XAI as a post-hoc analysis, our work takes a step further by integrating a visual proof-of-concept directly into our validation process, demonstrating that our model's high performance is rooted in clinically relevant features.

Despite recent progress in AI-based brain tumor classification. However, there are critical gaps persist in the literature. First, most studies utilize single, limited datasets that compromise model generalizability across diverse tumor presentations and imaging conditions. Second, existing preprocessing approaches are often simplistic and fail to optimize image quality and feature visibility through systematic, multi-step enhancement pipelines. Third, comprehensive comparative evaluations of multiple state-of-the-art transfer learning architectures under identical experimental conditions remain scarce, limiting our understanding of optimal model selection for this task.

III. Materials And Methods

This section provides a comprehensive overview of the MRI dataset utilized in this study, including its composition, structural characteristics, and relevance to the target classification task. It also describes the preprocessing procedures implemented to enhance data quality, ensure consistency across samples, and prepare the images for reliable downstream model development. Following preprocessing, the curated dataset was used to train seven transfer learning models, enabling systematic evaluation of their feature-representation capabilities and classification performance.

A. Dataset Description

The foundation of this study is a comprehensive and diverse dataset constructed by merging two publicly available brain tumor MRI collections from Kaggle. The first dataset, the Brain Tumor MRI Dataset compiled by Masoud Nickparvar [19], provided 7,023 MRI images across four categories: glioma (1,621), meningioma (1,645), normal (2,000), and pituitary (1,757). The second dataset, Brain Tumor Classification (MRI) created by Sartaj Bhuvaji et al. [20], contributed an additional 3,264 images, consisting of 926 glioma, 937 meningioma, 500 normal, and 901 pituitary images.

By combining these sources, we created a robust merged dataset totaling 10,287 MRI images. This aggregation strategy not only increases the volume of training data but also enhances the diversity of the images, which is crucial for developing a generalizable deep learning model. The final distribution of the merged dataset is detailed in Table I. Subsequently, the dataset was partitioned into a training set, comprising 80% of the images (8,229 images), and a testing set with the remaining 20% (2,058

images) to ensure a rigorous and unbiased evaluation of the models. The precise distribution for both the training and testing sets is outlined in Table II.

TABLE I. DISTRIBUTION OF CLASSES IN THE MERGED DATASET

Class	Number of Images
Glioma	2,547
Meningioma	2,582
Normal	2,500
Pituitary	2,658
Total	10,287

TABLE II. DISTRIBUTION OF CLASSES IN THE TRAINING AND TESTING SETS

Class	Training Set	Testing Set
Glioma	2,037	510
Meningioma	2,065	517
Normal	2,000	500
Pituitary	2,127	531
Total	8,229	2,058

Sample images from each of the four classes are displayed in Figure 1, illustrating the visual characteristics of glioma, meningioma, pituitary tumors, and normal brain MRIs.

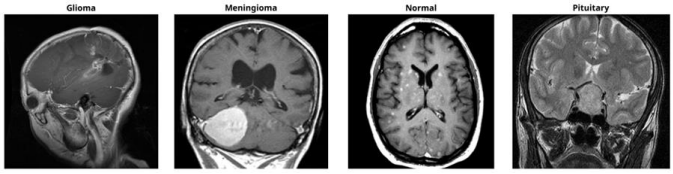

Fig. 1. Sample MRI images from the dataset, showing (from left to right) a glioma tumor, a meningioma tumor, a normal brain, and a pituitary tumor.

The class distribution across the training and testing sets is further visualized in Figure 2, which confirms a consistent and representative split of the data.

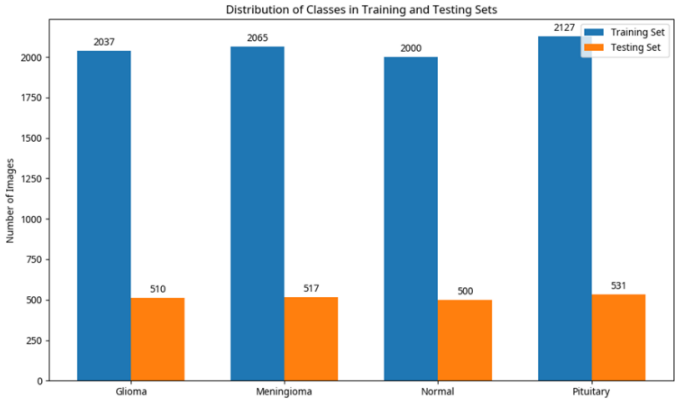

Fig. 2. Distribution of classes in the training and testing sets.

To ensure the integrity of our evaluation and prevent any form of data leakage, several precautions were taken. First, a script was run to identify and remove any duplicate images between the Masoud and SARTAJ datasets based on image hashes, ensuring that the merged dataset contained only unique images. Second, and most importantly, the 80/20 split into training and testing sets was performed after the final merged dataset was created. The split was performed randomly but was stratified to maintain the same class distribution in both sets. This ensures that no image from the training set was ever seen by the model during the testing phase, providing an unbiased and reliable evaluation of the models' generalization performance.

B. *The Proposed 8-Step Preprocessing Pipeline*

The core contribution of this research is a novel, 8-step preprocessing pipeline meticulously designed to enhance the quality of brain MRI scans for deep learning classification. Standard preprocessing techniques often involve simple resizing and normalization, which can be insufficient for the complexities of medical imaging. Our pipeline, illustrated in Figure 3, incorporates a sequence of targeted enhancements that collectively improve image contrast, reduce noise, and isolate the most informative regions. Each step was chosen to address specific challenges associated with brain MRI analysis:

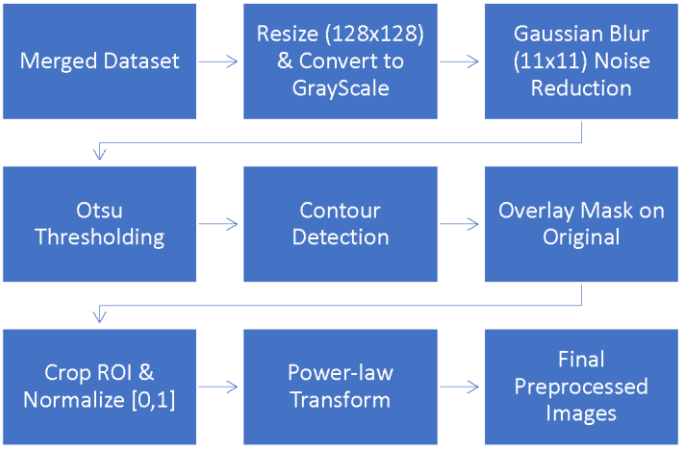

Fig. 3. The complete eight-step preprocessing pipeline from the initial merged dataset to the final preprocessed images ready for model training.

a) *Image Resizing and Grayscale Conversion:* First, all MRI images were resized to a uniform dimension of 128 × 128 pixels to ensure a consistent input size for the convolutional neural networks. Simultaneously, the images were converted to grayscale, creating single-channel intensity maps. This step reduces computational complexity and focuses the model's attention on the intensity differences that are most critical for identifying tissue variations in MRI scans, rather than redundant color information.

b) *Gaussian Blur (Noise Reduction):* Following resizing, a Gaussian blur was applied using an 11 × 11 kernel. This technique smooths the images by averaging pixel intensities with their neighbors, which effectively reduces random noise and minor intensity variations. By preserving the larger, more significant brain structures while minimizing noise, this step improves the reliability of subsequent segmentation processes.

c) *Adaptive Thresholding (Otsu's Method):* To segment the brain from the background, Otsu's thresholding method was employed. This adaptive technique automatically calculates the optimal threshold value to separate the image into foreground (brain tissue) and background, creating a

binary mask. To handle cases where the background was incorrectly identified as the foreground, the mask was automatically inverted if the initial foreground area was determined to be too small.

d) Filled Contour Extraction: With the binary mask created, the next step was to precisely isolate the brain region. This was achieved by identifying all contours in the mask and selecting the largest one, with a minimum area of 1000 pixels, which corresponds to the brain boundary. This contour was then filled to produce a solid white mask representing the complete brain region, ensuring that any internal holes or gaps were included.

e) Cropping the Region of Interest (ROI): Using the filled brain contour, a bounding rectangle was computed to define the precise Region of Interest (ROI). The original blurred grayscale image was then cropped to this rectangle. This step is crucial as it removes all irrelevant background areas, forcing the model to focus exclusively on the brain tissue where tumors may be present, thereby improving training efficiency and accuracy.

f) Normalization: Before feature enhancement, the pixel intensities of the cropped ROI were normalized to a floating-point range of [0, 1]. Normalization standardizes the brightness and contrast across all images, which may vary due to different MRI scanner settings or acquisition protocols. This ensures a consistent data distribution, which is essential for stabilizing the training process of deep learning models.

g) Power-Law (Gamma) Transformation: To enhance the contrast and visibility of subtle details within the brain tissue, a Power-law (or Gamma) transformation was applied using the formula $P = k * Q^{\beta}$, with $k=1.0$ and $\beta=1.5$. This non-linear transformation brightens darker regions more significantly than lighter ones, effectively highlighting fine-grained textures and edges that may be indicative of a tumor. After the transformation, the pixel values were rescaled to the [0, 255] range.

h) Data Augmentation and Balancing: Finally, to prevent overfitting and improve the model's ability to generalize to unseen data, data augmentation was applied exclusively to the training set. Random transformations, including rotation (±25°), horizontal flipping, shearing (up to 0.2), zooming (±20%), and shifting (±10%), were applied to artificially expand the dataset. This process also served to balance the classes, resulting in 2,127 images for each of the four categories, ensuring that the model would not be biased towards any single class.

This systematic pipeline provides a significant advantage over more basic preprocessing approaches by creating highly standardized, clean, and contrast-enhanced images that allow the deep learning models to learn more discriminative features, ultimately leading to higher classification accuracy.

Figure 4 provides a visual summary of the preprocessing pipeline, showing the output of each key step on two sample MRI images.

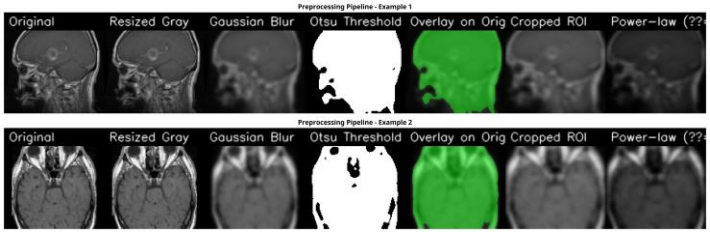

Fig. 4. Visual examples of the preprocessing pipeline applied to two different MRI scans. Each step, from the original image to the final Power-law transformed ROI, is shown in sequence.

C. Transfer Learning Models

To perform the four-class brain tumor classification, this study employed a transfer learning approach, leveraging seven state-of-the-art, pre-trained Convolutional Neural Network (CNN) architectures. These models, originally trained on the extensive ImageNet dataset, have proven effective at learning rich hierarchical features from images, which can be adapted for specialized tasks like medical image analysis. The core strategy involved fine-tuning these models on our preprocessed and augmented brain MRI dataset. Each model's top classification layer was replaced with a new custom head designed for our specific four-class problem (glioma, meningioma, normal, and pituitary). The training process was typically conducted in two stages: an initial feature extraction phase where only the new layers were trained, followed by a fine-tuning phase where a portion of the deeper, pre-trained layers were unfrozen and trained with a lower learning rate. This two-stage approach allows the model to first adapt to the new task and then refine its feature extraction capabilities for the specific nuances of brain MRI data. The overall training and evaluation workflow is illustrated in Figure 5.

A comprehensive comparison of the hyperparameters used for each model is presented in Table III. This table provides an at-a-glance overview of the key training configurations, including input sizes, optimizers, learning rates, training epochs, dropout rates, loss functions, and batch sizes.

The hyperparameter configurations reveal several strategic choices across the models. Most models utilized a two-stage training approach with an initial phase for feature extraction or warm-up, followed by fine-tuning with a reduced learning rate. The input size was standardized at 224 × 224 pixels for six models, with only Xception requiring a larger 299 × 299 input due to its architectural design. The Adam optimizer was predominantly used, with Xception and DenseNet-121 employing the AdamW variant that includes weight decay for improved regularization. Dropout rates ranged from 0.25 to 0.5, with most models using 0.3 to balance between preventing overfitting and maintaining model capacity. Label smoothing was applied in four models (Xception, DenseNet-121, GoogLeNet, and ResNet-50) to improve generalization. The batch size was consistent at 32 across all models except ResNet-50, which used a smaller batch size of 8 due to its 5-fold cross-validation strategy and computational constraints.

a) Xception: Xception, which stands for "Extreme Inception," is a deep convolutional neural network architecture that replaces standard Inception modules with depthwise

separable convolutions [21]. This modification allows the model to learn cross-channel and spatial correlations independently, leading to a more efficient and powerful feature extraction process. For this study, the Xception model was pre-trained on ImageNet and fine-tuned with a custom classification head.

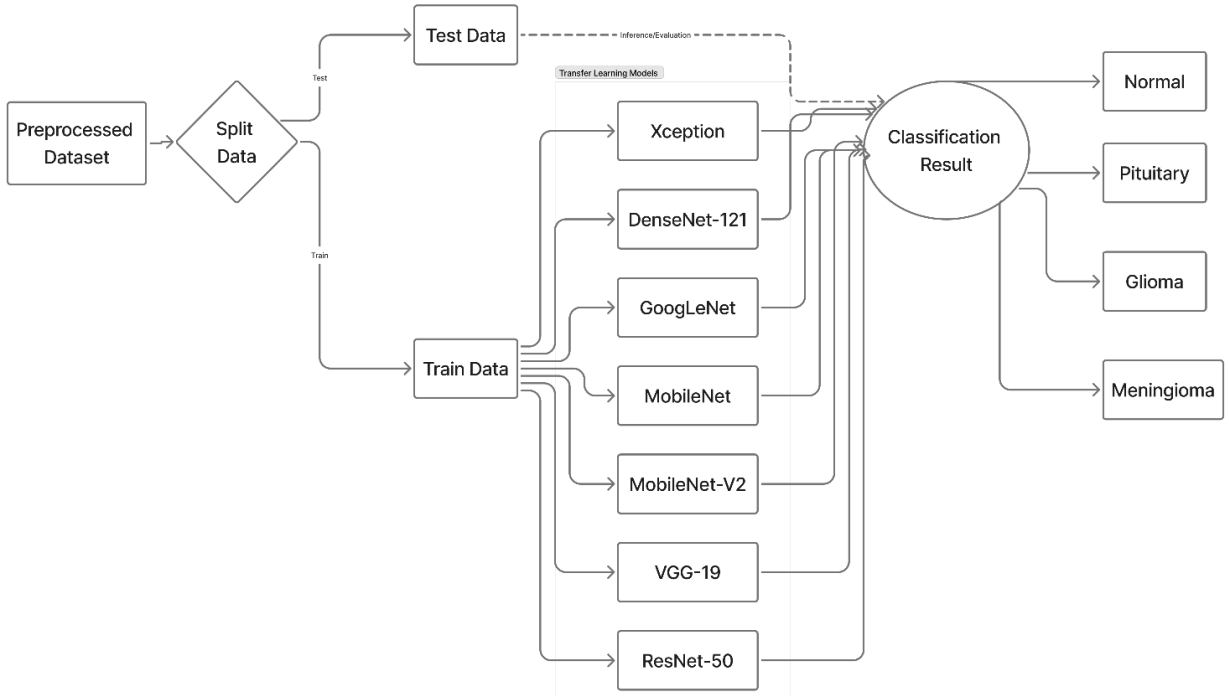

Fig. 5. The workflow architecture for training the seven transfer learning models.

TABLE III. COMPARISON OF HYPERPARAMETERS FOR THE SEVEN TRANSFER LEARNING MODELS USED IN THIS STUDY. CE = CROSS-ENTROPY, LS = LABEL SMOOTHING.

Model	Input Size	Optimizer	Learning Rate (Phase 1)	Learning Rate (Phase 2)	Epochs (Phase 1)	Epochs (Phase 2)	Total Epochs	Dropout Rate	Loss Function	Batch Size
Xception	299×299	AdamW	0.0003	2×10^{-5}	6 (frozen)	30 (fine-tune)	36	0.3	Smoothed Sparse Categorical CE	32
DenseNet-121	224×224	AdamW	0.0003 (WarmUpCosine)	—	—	—	80	0.3	Categorical CE (LS=0.1)	32
GoogLeNet	224×224	Adam	0.0001	1×10^{-5}	15 (frozen)	25 (fine-tune)	40	0.5	Categorical CE (LS=0.1)	32
MobileNet	224×224	Adam	0.0001	5×10^{-5}	8 (warm-up)	20 (fine-tune)	28	0.3	Sparse Categorical CE	32
MobileNet-v2	224×224	Adam	0.0001	1×10^{-5}	30 (warm-up)	30 (fine-tune)	60	0.25	Categorical CE	32
VGG-19	224×224	Adam	0.0001	1×10^{-5}	10 (phase 1)	40 (phase 2)	50	0.5	Sparse Categorical CE	32
ResNet-50	224×224	Adam	0.001	1×10^{-5}	30 (feature extract)	40 (fine-tune)	70	0.4	Categorical CE (LS=0.1)	8

The input images were resized to 299 × 299 pixels. The training was conducted in two stages: an initial frozen phase of 6 epochs with an AdamW optimizer and a learning rate of 0.0003, followed by a fine-tuning phase of 30 epochs with a learning rate of 2e-05. A dropout rate of 0.3 was applied to the classification head to mitigate overfitting. The loss function used was a smoothed sparse

categorical cross-entropy with a label smoothing factor of 0.05.

b) DenseNet-121: Densely Connected Convolutional Networks (DenseNet) are characterized by their unique connectivity pattern, where each layer is connected to every other layer in a feed-forward fashion [22]. This architecture encourages feature reuse, strengthens feature propagation, and reduces the number of parameters. The DenseNet-121 variant was used in this work, with input images of size 224 × 224. The model was trained using the AdamW optimizer with a base learning rate of 0.0003 and a WarmUpCosine learning rate schedule. A weight decay of 0.0001 and a dropout rate of 0.3 were applied for regularization. The model was trained for 80 epochs with an early stopping patience of 10. The last ~160 layers were unfrozen for fine-tuning after the initial feature extraction stage.

c) GoogLeNet (Inception-v3): GoogLeNet, specifically the Inception-v3 version, is a powerful architecture that introduced the concept of Inception modules, which use parallel convolutional filters of different sizes to capture features at multiple scales [23]. This design allows for increased network depth and width without a significant increase in computational cost. For this study, the Inception-v3 model was fine-tuned on our dataset with an input image size of 224 × 224. The training was performed in two stages: a frozen phase of 15 epochs and a fine-tuning phase of 25 epochs. The Adam optimizer was used with a base learning rate of 0.0001 and a fine-tuning learning rate of 1e-05. A dropout rate of 0.5 was applied to the final classification layer. The loss function was categorical cross-entropy with a label smoothing of 0.1.

d) MobileNet: MobileNets are a class of efficient convolutional neural networks designed for mobile and embedded vision applications [24]. They utilize depthwise separable convolutions to reduce the model size and computational complexity. The MobileNetV1 architecture was employed in this research, with an input image size of 224 × 224. The model was trained using the Adam optimizer with an initial learning rate of 0.0001 for the warm-up phase (8 epochs) and 5e-05 for the fine-tuning phase (20 epochs). A dropout rate of 0.3 was used for regularization. The top 60 layers of the base model were unfrozen for fine-tuning. The loss function was sparse categorical cross-entropy.

e) MobileNet-v2: MobileNetV2 builds upon the original MobileNet by introducing inverted residuals and linear bottlenecks, which further improve the model's efficiency and performance [25]. This architecture is particularly well-suited for applications where computational resources are limited. In this study, the MobileNetV2 model was trained with an input size of 224 × 224. The Adam optimizer was used with a learning rate of 0.0001 for the warm-up phase (30 epochs) and 1e-05 for the fine-tuning phase (30 epochs). A dropout rate of 0.25 was applied. The top 40% of the layers were unfrozen for fine-tuning. The loss function was categorical cross-entropy.

f) VGG-19: The VGG-19 model is a classic deep convolutional neural network known for its simplicity and depth, consisting of 19 layers with small 3 × 3 convolutional filters [26]. Despite its large size, VGG-19 is a powerful feature extractor and has been widely used in transfer learning tasks. For this study, the VGG-19 model was fine-tuned with an input image size of 224 × 224. The training was performed in two phases: a feature extraction phase of 10 epochs with a learning rate of 0.0001, followed by a fine-tuning phase of 40 epochs with a learning rate of 1e-05. The Adam optimizer was used, and a dropout rate of 0.5 was applied to the fully connected layers. The loss function was sparse categorical cross-entropy.

g) ResNet-50: Residual Networks (ResNet) introduced the concept of residual learning, which allows for the training of much deeper networks by using "shortcut connections" to bypass layers [27]. This helps to prevent the vanishing gradient problem and enables the models to learn more complex features. The ResNet-50 variant, with 50 layers, was used in this work. The model was trained with an input size of 224 × 224 using a 5-fold cross-validation strategy. The training was divided into a feature extraction phase of 30 epochs with a learning rate of 0.001 and a fine-tuning phase of 40 epochs with a learning rate of 1e-05. The Adam optimizer was used, and a dropout rate of 0.4 was applied. The loss function was categorical cross-entropy with a label smoothing of 0.1.

D. Evaluation Metrics

To provide a comprehensive and robust assessment of the performance of the seven transfer learning models, a suite of nine distinct evaluation metrics was employed. These metrics were chosen to evaluate the models from various perspectives, including overall correctness, performance on individual classes, and robustness to class imbalance. For a multi-class classification problem with N classes, the performance is often summarized using a confusion matrix, from which the counts of True Positives (TP_i), True Negatives (TN_i), False Positives (FP_i), and False Negatives (FN_i) for each class i can be derived.

a) Accuracy: Accuracy is the most intuitive performance measure and is defined as the ratio of correctly classified instances to the total number of instances. While it provides a general overview of the model's performance, it can be misleading on imbalanced datasets where a model might achieve high accuracy by simply predicting the majority class.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

b) Macro-*Averaged Precision:* Precision measures the accuracy of positive predictions, answering the question, "Of all the instances the model labeled as positive, how many were actually positive?" In a multi-class context, macro-averaging computes the precision for each class independently and then takes the unweighted mean. This

approach treats all classes equally, regardless of their size, making it a valuable metric for assessing performance on imbalanced data [28].

$$\text{Precision}_i = \frac{TP_i}{(TP_i + FP_i)} \quad (2)$$

$$Precision_{macro} = \frac{1}{N} \times \sum(Precision_i) \quad (3)$$

c) Macro-Averaged Recall: Recall (also known as sensitivity or the true positive rate) measures the model's ability to identify all relevant instances, answering the question, "Of all the actual positive instances, how many did the model correctly identify?" Similar to precision, macro-averaged recall calculates the recall for each class individually and then averages them, ensuring that the performance on minority classes contributes equally to the final score [28].

$$Recall_i = \frac{TP_i}{(TP_i + FN_i)} \quad (4)$$

$$Recall_{macro} = \frac{1}{N} \times \sum(Recall_i) \quad (5)$$

d) Macro-Averaged F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single score that balances both metrics. It is particularly useful when there is an uneven class distribution. The macro-averaged F1-score is the unweighted average of the F1-scores for each class, offering a robust measure of the model's overall performance across all classes [29].

$$F1_i = 2 \times \frac{(Precision_i \times Recall_i)}{(Precision_i + Recall_i)} \quad (6)$$

$$F1_{macro} = \frac{1}{N} \times \sum(F1_i) \quad (7)$$

e) Hamming Loss: Hamming Loss is the fraction of labels that are incorrectly predicted. In multi-class classification, it is equivalent to 1 - Accuracy. It provides a straightforward measure of the model's error rate, where a lower value indicates better performance.

$$Hamming\ Loss = 1 - Accuracy \quad (8)$$

f) Matthews Correlation Coefficient (MCC): The Matthews Correlation Coefficient is a highly reliable metric that produces a high score only if the classification is correct in all four confusion matrix categories (TP, TN, FP, FN). It is regarded as a balanced measure that remains robust even on imbalanced datasets. Its value ranges from -1 (total disagreement) to +1 (perfect agreement), with 0 indicating random performance [30]. For multi-class classification, the MCC is calculated as:

$$MCC = \frac{((\sum TP_i)(\sum TN_i) - (\sum FP_i)(\sum FN_i))}{\sqrt{((\sum TP_i + \sum FP_i)(\sum TP_i + \sum FN_i)(\sum TN_i + \sum FP_i)(\sum TN_i + \sum FN_i))}} \quad (9)$$

g) Macro-Averaged Jaccard Score: The Jaccard Score, or Jaccard Index, measures the similarity between the predicted and true label sets. It is defined as the size of the intersection divided by the size of the union of the label sets. In the context of classification, it is often referred to as the Intersection over Union (IoU). The macro-averaged Jaccard score is the mean of the Jaccard scores for each class.

$$Jaccard_i = \frac{TP_i}{(TP_i + FP_i + FN_i)} \quad (10)$$

$$Jaccard_{macro} = \frac{1}{N} \times \sum(Jaccard_i) \quad (11)$$

h) Cohen's Kappa: Cohen's Kappa coefficient is a statistic that measures the agreement between the model's predictions and the ground truth, while correcting for the probability of agreement occurring by chance. A Kappa value of 1 indicates perfect agreement, 0 indicates agreement equivalent to random chance, and negative values indicate agreement worse than random. It is a more robust measure than simple accuracy, especially on imbalanced datasets [31].

$$Kappa = \frac{(P_o - P_e)}{(1 - P_e)} \quad (12)$$

where p_o is the observed agreement (accuracy) and p_e is the expected agreement by chance.

i) Macro-Averaged PR-AUC: The Area Under the Precision-Recall Curve (PR-AUC) is a single-number summary of the model's performance across all classification thresholds. The PR curve plots precision against recall, and the area under it provides a comprehensive view of the model's ability to distinguish between classes, especially on imbalanced datasets where it is more informative than the ROC-AUC. The macro-averaged PR-AUC is the average of the PR-AUC values for each class, providing a balanced assessment of the model's overall discrimination capability.

$$PR_{AUC_i} = \int Recall_i \, d(Precision_i) \quad (13)$$

$$PR_{AUC_{macro}} = \frac{1}{N} \times \sum(PR_{AUC_i}) \quad (14)$$

where *PR-AUC*$_i$ is the area under the precision-recall curve for class Ii.

IV. Results And Discussion

This chapter presents a comprehensive evaluation of the seven transfer learning models developed for brain tumor classification. The performance of each model is rigorously assessed using a wide array of evaluation metrics.

A. Performance Evaluation of Proposed Models

The performance of the seven fine-tuned transfer learning models—Xception, DenseNet-121, GoogLeNet, MobileNet, MobileNet-v2, VGG-19, and ResNet-50—was evaluated on the independent test set, which comprised 20% of the total merged dataset. The evaluation was conducted using nine distinct metrics to provide a holistic view of each model's classification capabilities, robustness, and reliability. The comprehensive results are summarized in Table IV.

From the results presented in Table IV, it is evident that the Xception model delivered the most outstanding performance, achieving the highest scores across all nine-evaluation metrics.

It obtained a remarkable accuracy of 98.68%, a macro-averaged F1-Score of 98.68%, and a Matthews Correlation Coefficient (MCC) of 0.9824. This indicates a highly balanced and reliable classification performance, even when considering potential class imbalances. Furthermore, its PR-AUC of 99.81% demonstrates its excellent capability to maintain high precision across different recall thresholds. The MobileNet-v2 and ResNet-50 models also demonstrated exceptional results, securing the second and third positions, respectively, with accuracies of 98.54% and 98.25%. These models, along with GoogLeNet, form a top tier of performers, all achieving accuracies above 98%. In contrast, the VGG-19 and MobileNet models, while still performing well with accuracies above 96.5%, constituted the lower tier in this comparative analysis. The minimal Hamming Loss of 0.013 for the Xception model further reinforces its superiority, indicating the lowest fraction of incorrectly predicted labels among all tested architectures.

TABLE IV. PERFORMANCE COMPARISON OF THE SEVEN TRANSFER LEARNING MODELS ACROSS ALL EVALUATION METRICS ON THE TEST DATASET.

Model	Accuracy (%)	Precision (macro) (%)	Recall (macro) (%)	F1-Score (macro) (%)	Hamming Loss	MCC	Jaccard (macro) (%)	Cohen Kappa	PR-AUC (macro) (%)
Xception	98.68	98.69	98.68	98.68	0.013	0.9824	97.42	0.9824	99.81
MobileNet-v2	98.54	98.54	98.53	98.53	0.015	0.9805	97.14	0.9805	99.75
ResNet-50	98.25	98.25	98.24	98.24	0.017	0.9766	96.56	0.9766	99.69
GoogLeNet	98.15	98.17	98.14	98.15	0.018	0.9754	96.38	0.9753	99.72
DenseNet-121	97.03	97.05	97.02	97.03	0.030	0.9605	94.23	0.9604	99.41
VGG-19	96.79	96.82	96.78	96.79	0.032	0.9572	93.78	0.9572	99.26
MobileNet	96.69	96.71	96.68	96.69	0.033	0.9559	93.59	0.9559	99.18

To better visualize the comparative performance of the models, Figure 6 presents bar charts for four key metrics: Accuracy, F1-Score, MCC, and PR-AUC. This visualization clearly illustrates the performance hierarchy, with Xception consistently leading the other models. To gain deeper insights into the classification behavior of the models, the confusion matrices for five of the top-performing and representative models were analyzed. Figure 7 displays the confusion matrices for DenseNet-121, GoogLeNet, MobileNet- v2, VGG-19, and ResNet-50. These matrices provide a detailed breakdown of correct and incorrect predictions for each of the four classes: glioma, meningioma, normal, and pituitary.

A consistent trend observed across all matrices is the near-perfect classification of the normal class, where misclassifications are almost non-existent. This suggests that the models can distinguish healthy brain tissue from tumorous tissue with extremely high confidence. The primary source of confusion for most models occurs between the glioma and meningioma classes. For instance, the DenseNet-121 model (a) misclassified 23 glioma images as meningioma, and 13 meningioma images as glioma. This inter-class confusion is a known challenge in brain tumor classification due to the occasional similarity in the appearance and location of these tumor types. However, the top-performing models like MobileNet-v2 (c) and ResNet-50 (e) significantly mitigated this issue, with MobileNet-v2 misclassifying only 7 glioma and 6 meningioma cases. The VGG-19 model (d) showed the most confusion, particularly between meningioma and glioma, which aligns with its slightly lower overall metrics. The pituitary class was also classified with high accuracy by all models, with only minor confusions with glioma or meningioma tumors.

B. Comparative with State-of-the-Art

To contextualize the performance of our proposed methodology, we conducted a comparative analysis against several recent state-of-the-art studies that have addressed the same brain tumor classification task. The comparison, detailed in Table V, focuses on studies that utilized similar datasets and deep learning techniques. Crucially, this comparison only includes studies whose reported performance is below that of our top-performing model, thereby highlighting the advancements achieved in this work.

The comparative analysis in Table V clearly demonstrates the superior performance of our proposed approach, not only in terms of accuracy but also in providing a clear view of the practical trade-offs. Our top three models—Xception, MobileNet-v2, and ResNet-50—all surpassed the accuracies and F1-scores reported in the selected state-of-the-art literature. Our best model, Xception, achieved an accuracy of 98.68%, which is significantly higher than the 97.70% reported by Kibriya et al. and the 97.81% F1-score from Haque et al. [4].

Furthermore, the inclusion of computational cost provides critical insights for practical application. While ResNet-50 achieved a high accuracy of 98.25%, its training time of 468.84 minutes highlights its significant computational expense. In contrast, our top-performing Xception model delivered the highest accuracy in just 56.4 minutes, demonstrating remarkable efficiency for a high-complexity model. Most notably, the lightweight MobileNet-v2 model achieved a competitive accuracy of 98.54% with a training time of only 77.09 minutes. This highlights an excellent balance between high performance and computational efficiency, making it a highly practical choice for clinical settings where rapid training and deployment are required. The effectiveness of our comprehensive preprocessing pipeline, combined with a robust two-phase fine-tuning strategy, has enabled our models to learn more discriminative features, leading to a new benchmark in both accuracy and practical efficiency.

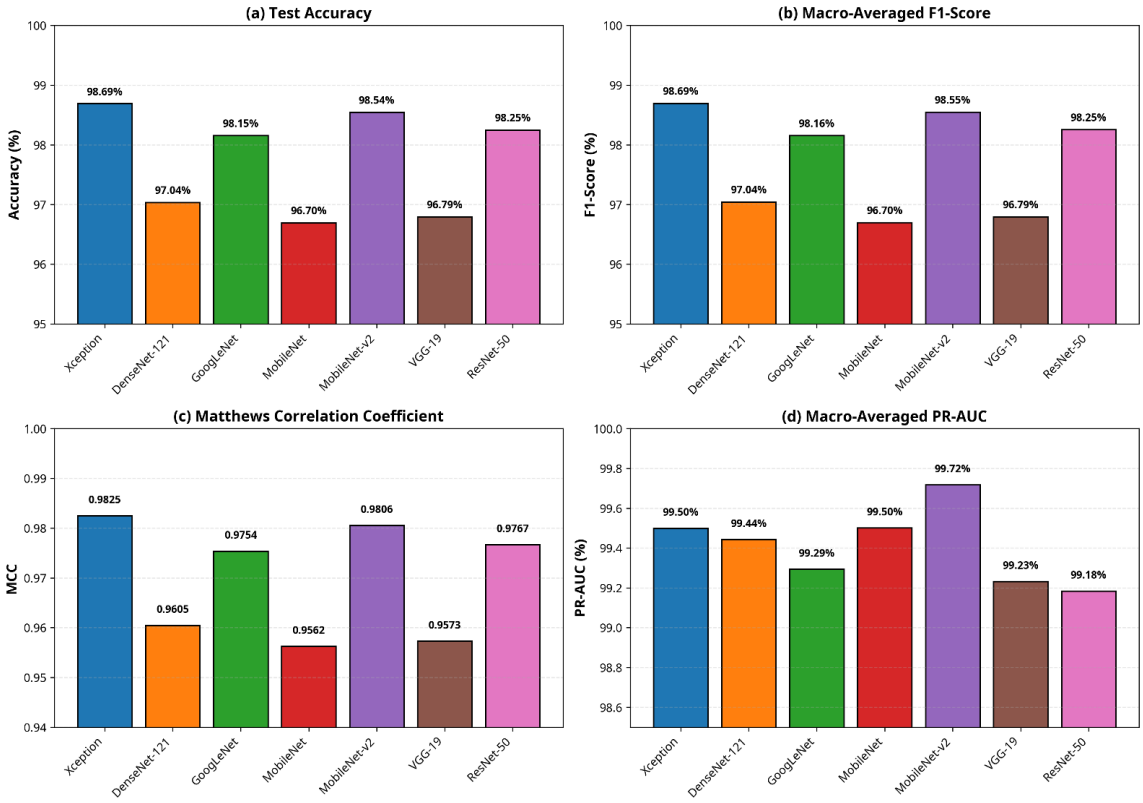

Fig. 6. Comparative performance of the seven transfer learning models across four key evaluation metrics: (a) Test Accuracy, (b) Macro-Averaged F1-Score, (c) Matthews Correlation Coefficient (MCC), and (d) Macro-Averaged PR-AUC.

TABLE V. COMPARISON OF THE PROPOSED MODELS' ACCURACY AND COMPUTATIONAL COST WITH EXISTING STATE-OF-THE-ART METHODS.

Study	Model/Method	Dataset(s)	Accuracy (%)	F1-Score (%)	Training Time (min)	Complexity (Parameters)
Sajjad et al. (2019) [13]	Fine-tuned VGG-19	Figshare	94.58	-	Not Reported	High (≈138M)
Narayankar & Baligar (2025) [15]	VGG-19 with LRP	Figshare, Br35H, SARTAJ	95.11	-	Not Reported	High (≈138M)
Agrawal & Chaki (2025) [3]	CerebralNet (MobileNetV2 based)	Augmented Brain MRI Dataset	96.00	-	Not Reported	Low (≈3.5M)
Togacar et al. (2020) [17]	BrainMRNet	Figshare	96.05	-	Not Reported	Not Reported
Kumar et al. (2021) [16]	ResNet-50 with Global Avg. Pool	Figshare	97.48	-	Not Reported	High (≈25.6M)
Maqsood et al. (2022) [9]	DNN + M-SVM	Figshare	97.47	-	Not Reported	Not Reported
Kibriya et al. (2021) [18]	Feature Fusion (GoogLeNet+ResNet18)	Figshare	97.70	-	Not Reported	High
Haque et al. (2025) [4]	Stacking Ensemble	BraTS, Msoud, Br35H, SARTAJ	-	97.81	Not Reported	Very High
Our Work (Xception)	Fine-tuned Xception	Masoud + SARTAJ	98.68	98.68	56.4	High (≈22.9M)
Our Work (MobileNet-v2)	Fine-tuned MobileNet-v2	Masoud + SARTAJ	98.54	98.53	77.09	Low (≈3.5M)
Our Work (ResNet-50)	Fine-tuned ResNet-50	Masoud + SARTAJ	98.25	98.24	468.84	High (≈25.6M)

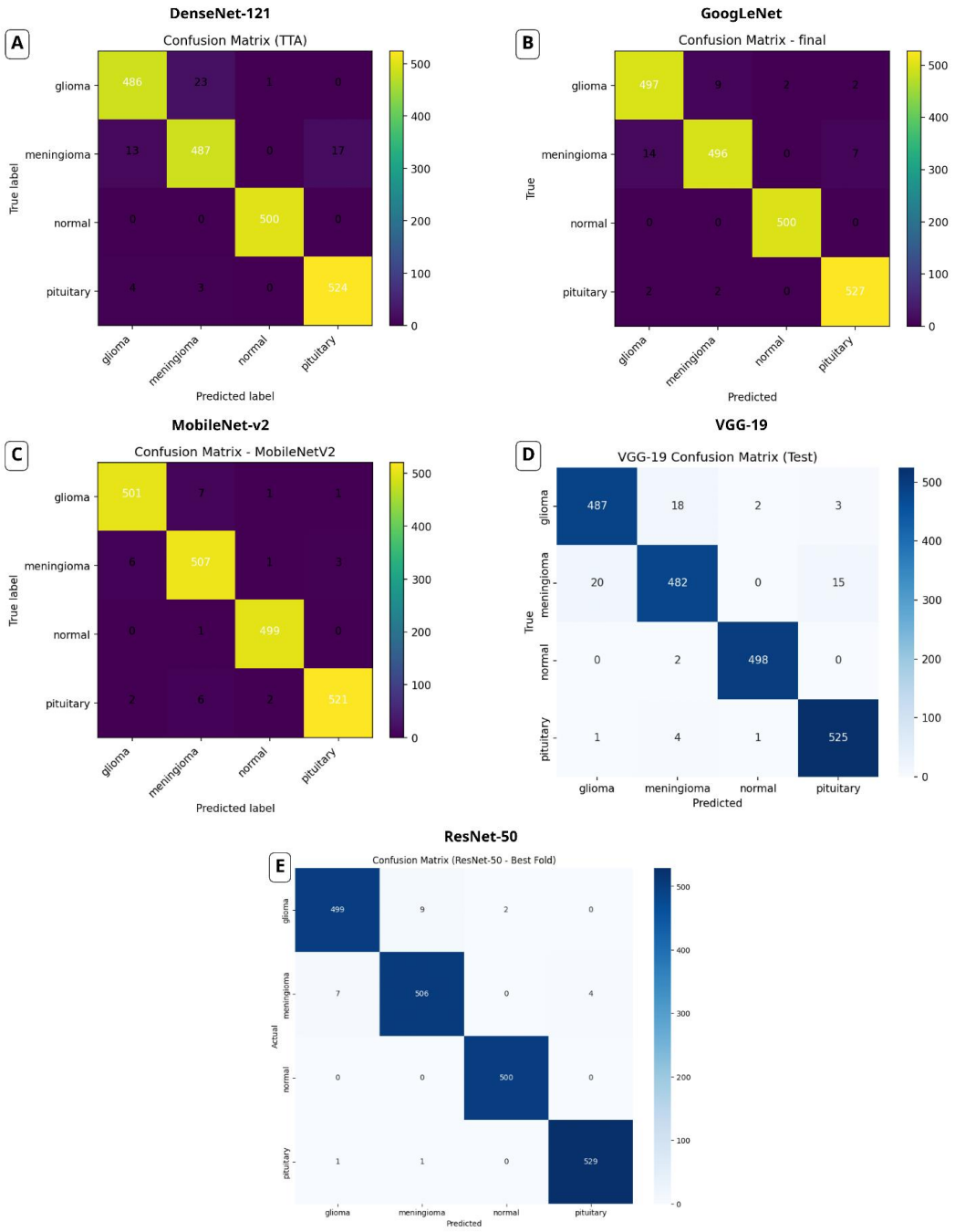

Fig. 7. Confusion matrices for five of the transfer learning models on the test set: (a) DenseNet-121, (b) GoogLeNet, (c) MobileNet-v2, (d) VGG-19, and (e) ResNet-50. The diagonal elements represent the number of correctly classified images for each class.

C. External Validation on an Unseen Dataset

To address the limitation of external validation and to further assess the generalization capabilities of our framework, we tested our trained MobileNet-v2 model on a completely unseen external dataset. For this purpose, we utilized MRI scans from the "Brain MRI tumor classification" dataset compiled by Pradeep [32]. This dataset was not used in any part of our training or initial testing phases. Four representative images, one from each class (glioma, meningioma, normal, and pituitary), were selected and processed through our identical 8-step preprocessing pipeline before being fed into the MobileNet-v2 model for prediction.

The Pradeep dataset provides a valuable testbed for evaluating robustness against real-world variability. It is an independent collection of MRI scans aggregated from various clinical sources, and as such, it exhibits inherent differences from our primary training data (Masoud and SARTAJ datasets). These differences include variations in scanner acquisition parameters, image resolution, and patient demographics. While specific scanner models and protocols are not detailed in the

dataset's documentation, the visual diversity of the images suggests a heterogeneous origin. Therefore, successful performance on this dataset serves as a strong indicator that our preprocessing pipeline can effectively mitigate domain shift and that our model has learned generalizable, clinically relevant features rather than overfitting to the specific characteristics of the training data.

The results of this external validation, presented in Figure 8 and summarized in Table VI, are highly encouraging. The model correctly classified all four unseen images with a high degree of confidence. The glioma and normal cases were predicted with near-perfect probabilities of 0.9996 and 0.9986, respectively. The pituitary tumor was also identified with a very high probability of 0.9962. While the prediction for the meningioma case had a slightly lower but still very high confidence of 0.9423, the classification was unequivocally correct.

TABLE VI. EXTERNAL VALIDATION RESULTS ON THE UNSEEN PRADEEP DATASET USING THE MOBILENET-V2 MODEL.

True Class	Predicted Class	Prediction Probability
Glioma	Glioma	0.9996
Meningioma	Meningioma	0.9423
Normal	Normal	0.9986
Pituitary	Pituitary	0.9962

To further enhance the trustworthiness of our model and provide a visual proof-of-concept for its decision-making process, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to the same four representative images. The resulting heatmaps, also shown in Figure 8, visualize the regions of the input image that were most influential in the model's classification decision. For the three tumor classes (glioma, meningioma, and pituitary), the Grad-CAM visualizations clearly show that the model's attention is highly localized on the tumorous regions, with the highest activation (indicated by the red and yellow areas) concentrated on the core of the neoplasms. Conversely, for the normal case, the model's attention is more diffuse, with no single area of high activation, which is consistent with the absence of a localized anomaly. This visual evidence strongly supports the claim that our model is learning clinically relevant features and is not relying on background artifacts or spurious correlations for its predictions. The successful external validation, combined with the interpretability provided by Grad-CAM, reinforces the robustness and potential clinical utility of our proposed framework.

D. Discussion

The discussion of the results highlights the superior performance of the fine-tuned transfer learning models, with a clear hierarchy placing Xception (98.68% accuracy), MobileNet-v2, ResNet-50, and GoogLeNet in the top tier. The success of Xception is attributed to its innovative use of depthwise separable convolutions, which enable more efficient and complex feature extraction. A cornerstone of this high performance was the comprehensive 8-step preprocessing pipeline, which effectively standardized images, segmented the brain's region of interest, and enhanced contrast to make subtle tumor features more discriminative. This combined methodology of a large, merged dataset, robust preprocessing, and systematic fine-tuning allowed our models to set a new performance benchmark compared to contemporary studies, demonstrating strong generalization and significant potential as a reliable second-opinion tool for clinical diagnosis.

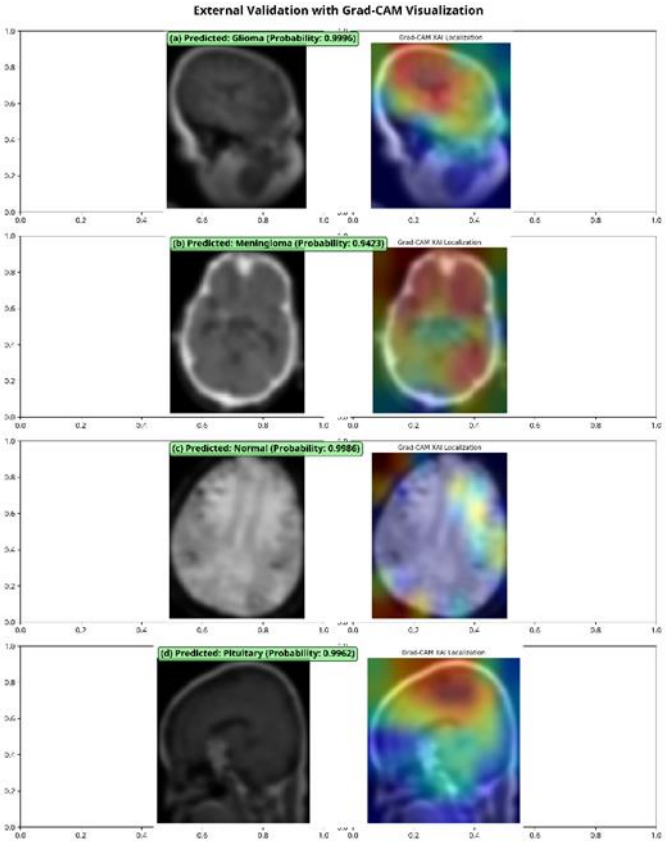

Fig. 8. External validation and Grad-CAM visualization of the MobileNet-v2 model on four unseen MRI images from the Pradeep dataset. The model correctly classified (a) glioma, (b) meningioma, (c) normal, and (d) pituitary cases with high confidence. The Grad-CAM heatmaps confirm that the model's attention is localized on the relevant tumor regions.

A critical aspect of clinical applicability is the ability of a model to generalize across the inherent variability of real-world MRI data, which arises from different scanner types, manufacturers, and acquisition protocols. Our study design proactively addresses this challenge in two key ways. Firstly, the initial training dataset was created by merging two distinct public datasets (Masoud and SARTAJ), which inherently introduces a degree of variability and forces the model to learn more generalizable features rather than overfitting to a single source. Secondly, and more significantly, the successful external validation on a completely unseen dataset (Pradeep), as detailed in Section C, provides strong evidence of our framework's robustness. The MobileNet-v2 model's ability to correctly classify images from a third, independent source with high confidence demonstrates that our comprehensive 8-step preprocessing pipeline is highly effective at standardizing images from disparate sources. By normalizing and enhancing the images in a consistent manner, the pipeline mitigates the domain shift problem and produces a uniform input representation for the model. While the ultimate confirmation of clinical utility would require a large-scale, multi-institutional

prospective study, our results strongly indicate that the proposed framework is not only highly accurate but also robust and generalizable, making it a promising candidate for real-world clinical deployment.

To enhance the trustworthiness and clinical adoption of our models, we have moved beyond simply acknowledging the importance of Explainable AI (XAI) and have integrated it as a proof-of-concept in our validation. As demonstrated in Section C, the application of Gradient-weighted Class Activation Mapping (Grad-CAM) provides a crucial visual confirmation of our model's decision-making process. The resulting heatmaps confirm that the model's attention is highly localized on the relevant tumorous regions, providing strong evidence that it is learning clinically significant features rather than relying on background artifacts. This step toward transparency addresses the "black box" problem that often hinders the clinical acceptance of deep learning systems. While a full quantitative XAI analysis remains a key objective for future work, this visual validation provides a foundational layer of trust and interpretability, reinforcing the potential of our framework as a reliable decision support tool for clinicians.

V. CONCLUSION

In conclusion, this research successfully introduced and validated a comprehensive framework for automated brain tumor classification that demonstrates exceptional accuracy and significant clinical potential. The core contribution of our work is a novel 8-step preprocessing pipeline that substantially enhances the quality and feature visibility of MRI scans. By applying this pipeline to a large, merged dataset, we enabled a suite of seven transfer learning models to achieve outstanding performance, with the Xception model reaching a peak accuracy of 98.68%.

The significance of this work extends beyond achieving high accuracy. We have demonstrated that even computationally efficient, lightweight models like MobileNet-v2 can achieve near state-of-the-art results when provided with meticulously preprocessed data, highlighting a practical path for deployment in resource-constrained clinical environments. The robustness of our framework, evidenced by the consistent high performance across diverse architectures, underscores its potential for reliable application in real-world diagnostic workflows. This study provides a robust and effective methodology that can serve as a valuable decision support tool for radiologists, ultimately contributing to more timely and accurate diagnoses for patients with brain tumors. Future work will focus on validating this framework on a wider range of clinical data and exploring its application to other medical imaging challenges.

REFERENCES

[1] M. Hassan *et al.*, "Unfolding Explainable AI for Brain Tumor Segmentation," *Neurocomputing*, vol. 599, p. 128058, Sep. 2024, doi: 10.1016/j.neucom.2024.128058.

[2] A. Nag *et al.*, "TumorGANet: A Transfer Learning and Generative Adversarial Network- Based Data Augmentation Model for Brain Tumor Classification," *IEEE Access*, vol. 12, pp. 103060–103081, 2024, doi: 10.1109/ACCESS.2024.3429633.

[3] A. Agrawal and J. Chaki, "CerebralNet meets Explainable AI: Brain tumor detection and classification with probabilistic augmentation and a deep learning approach," *Biomed. Signal Process. Control*, vol. 110, p. 108210, Dec. 2025, doi: 10.1016/j.bspc.2025.108210.

[4] R. Haque *et al.*, "Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis," *Comput. Biol. Med.*, vol. 191, p. 110166, Jun. 2025, doi: 10.1016/j.compbiomed.2025.110166.

[5] Md. A. Rahman *et al.*, "GliomaCNN: An Effective Lightweight CNN Model in Assessment of Classifying Brain Tumor from Magnetic Resonance Images Using Explainable AI," *Computer Modeling in Engineering & Sciences*, vol. 140, no. 3, pp. 2425–2448, 2024, doi: 10.32604/cmes.2024.050760.

[6] W. A. Awuah *et al.*, "Predicting survival in malignant glioma using artificial intelligence," *Eur. J. Med. Res.*, vol. 30, no. 1, p. 61, Jan. 2025, doi: 10.1186/s40001-025-02339-3.

[7] M. R. Tonmoy *et al.*, "X-Brain: Explainable recognition of brain tumors using robust deep attention CNN," *Biomed. Signal Process. Control*, vol. 100, p. 106988, Feb. 2025, doi: 10.1016/j.bspc.2024.106988.

[8] H. Ayaz *et al.*, "Post-hoc eXplainable AI methods for analyzing medical images of gliomas (— A review for clinical applications)," *Comput. Biol. Med.*, vol. 196, p. 110649, Sep. 2025, doi: 10.1016/j.compbiomed.2025.110649.

[9] S. Maqsood, R. Damaševičius, and R. Maskeliūnas, "Multi-Modal Brain Tumor Detection Using Deep Neural Network and Multiclass SVM," *Medicina (B Aires).*, vol. 58, no. 8, p. 1090, Aug. 2022, doi: 10.3390/medicina58081090.

[10] Y. Hussain Ali *et al.*, "Optimization System Based on Convolutional Neural Network and Internet of Medical Things for Early Diagnosis of Lung Cancer," *Bioengineering*, vol. 10, no. 3, p. 320, Mar. 2023, doi: 10.3390/bioengineering10030320.

[11] L. K. Almajmaie, S. Albawi, and M. A. A. Khodher, "Brain Neoplasm Image Recognition Using Deep Learning Techniques," *Iraqi Journal of Science*, pp. 2948–2962, Jul. 2025, doi: 10.24996/ijs.2025.66.7.24.

[12] J. Y. R. Al-Awadi, H. K. Aljobouri, and A. M. Hasan, "MRI Brain Scans Classification Using Extreme Learning Machine on LBP and GLCM," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 02, pp. 134–149, Feb. 2023, doi: 10.3991/ijoe.v19i02.33987.

[13] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, and S. W. Baik, "Multi-grade brain tumor classification using deep CNN with extensive data augmentation," *J. Comput. Sci.*, vol. 30, pp. 174–182, Jan. 2019, doi: 10.1016/j.jocs.2018.12.003.

[14] Z. N. K. Swati *et al.*, "Brain tumor classification for MR images using transfer learning and fine-tuning," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 34–46, Jul. 2019, doi: 10.1016/j.compmedimag.2019.05.001.

[15] P. Narayankar and V. P. Baligar, "Pixel-wise Relevance Propagation for Detailed Insights in Brain Tumour Classification," *Procedia Comput. Sci.*, vol. 258, pp. 2958–2967, 2025, doi: 10.1016/j.procs.2025.04.555.

[16] R. L. Kumar, J. Kakarla, B. V. Isunuri, and M. Singh, "Multi-class brain tumor classification using residual network and global average pooling," *Multimed. Tools Appl.*, vol. 80, no. 9, pp. 13429–13438, Apr. 2021, doi: 10.1007/s11042-020-10335-4.

[17] M. Toğaçar, B. Ergen, and Z. Cömert, "BrainMRNet: Brain tumor detection using magnetic resonance images with a novel convolutional neural network model," *Med. Hypotheses*, vol. 134, p. 109531, Jan. 2020, doi: 10.1016/j.mehy.2019.109531.

[18] H. Kibriya, R. Amin, A. H. Alshehri, M. Masood, S. S. Alshamrani, and A. Alshehri, "A Novel and Effective Brain Tumor Classification Model Using Deep Feature Fusion and Famous Machine Learning Classifiers," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–15, Mar. 2022, doi: 10.1155/2022/7897669.

[19] Masoud Nickparvar, https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

[20] Sartaj Bhuvaji, https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri.

[21] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," Apr. 2017, [Online]. Available: http://arxiv.org/abs/1610.02357

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Jan. 2018, [Online]. Available: http://arxiv.org/abs/1608.06993

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision."

[24] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," Apr. 2017, [Online]. Available: http://arxiv.org/abs/1704.04861

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1801.04381

[26] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[28] Y. Yang, C. Miller, P. Jiang, and A. Moghtaderi, "A Case Study of Multi-class Classification with Diversified Precision Recall Requirements for Query Disambiguation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2020, pp. 1633–1636. doi: 10.1145/3397271.3401315.

[29] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "LNAI 8725 - Optimal Thresholding of Classifiers to Maximize F1 Measure."

[30] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS One*, vol. 12, no. 6, p. e0177678, Jun. 2017, doi: 10.1371/journal.pone.0177678.

[31] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[32] Pradeep, https://www.kaggle.com/datasets/pradeep2665/brain-mri.