Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Comparative Study of Deep Learning Models for MRI-based Brain Tumor Classification

Muhammad Naufal Erza Farandi [1,*], Azah Kamilah Muda [2], Sri Winarno [1], and Halizah Basiron [2]

[1] Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang 50131, Indonesia; e-mail : erza.naufal@gmail.com; sri.winarno@dsn.dinus.ac.id

[2] Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia Melaka, Malacca 76100, Malaysia; e-mail : azah@utem.edu.my; halizah@utem.edu.my

* Corresponding Author : Muhammad Naufal Erza Farandi

Abstract: Brain tumor detection using deep learning has become a focus of research due to its potential to improve patient diagnosis and management. This study presents a comparative analysis to evaluate how the fundamental design philosophies of convolutional neural network (CNN) architectures, namely sequential (VGG16), residual (ResNet50), and encoder-decoder (U-Net), influence performance on brain tumor classification tasks using a public MRI dataset. The data undergoes preprocessing, including normalization, data augmentation, and division into training and testing subsets. Evaluation is conducted using accuracy, precision, recall, and F1-Score metrics. The results show that ResNet50 consistently outperforms other architectures with an average accuracy of 95.35%, followed by VGG16 (93.93%). Conversely, U-Net, designed for segmentation, demonstrates the lowest performance (79%). The superiority of ResNet50 highlights the benefits of residual connections in addressing the vanishing gradient problem for classification tasks. The poor performance of U-Net confirms the hypothesis that architectural suitability for a specific task is more critical than complexity alone. This study highlights the strength of ResNet50 as a reliable approach for brain tumor classification. It underscores the importance of selecting an architecture specifically designed for the intended medical image analysis task. In addition, an analysis of model complexity and computational efficiency was conducted, showing that while ResNet50 achieved the highest accuracy, VGG16 provided a more favorable trade-off between performance and training time.

Keywords: Brain Tumor Classification; Convolutional Neural Networks; Deep Learning; Magnetic Resonance Imaging; Medical Image Analysis; Model Complexity; Residual Connections; U-Net.

Received: August, 7th 2025
Revised: September, 7th 20205
Accepted: September, 10th 2025
Published: September, 15th 2025
Curr. Ver.: September, 15th 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Brain tumors are abnormal cell growths within or around brain tissue, which occur when cells undergo uncontrolled division and deviate from the normal cell cycle. According to the World Health Organization (WHO), brain tumors account for about 2% of all cancer cases and have a high mortality rate, especially for malignant tumors [1], [2]. This condition is hazardous due to several critical factors: the brain is a vital organ that controls all bodily functions, and tumor growth within the confined intracranial space can rapidly cause dangerous pressure that disrupts neurological function [3]. Additionally, the blood-brain barrier, which protects the brain, can also hinder the effectiveness of treatment [4], [5].

For diagnostic and treatment purposes, brain tumors are classified based on several criteria. The main distinction is made between benign tumors, which grow slowly with clear boundaries and a low likelihood of recurrence, and malignant tumors, which exhibit rapid growth, unclear boundaries, and a tendency to invade surrounding tissues [6]. The WHO has developed a comprehensive histological classification system, grouping tumors into main categories such as glioma, meningioma, and pituitary adenoma. This system also includes an assessment of malignancy levels from Grade I (least aggressive with a good prognosis) to

Grade IV (most aggressive with a poor prognosis), which is an important reference in determining a patient's treatment plan [7], [8].

The diagnosis of brain tumors relies heavily on modern imaging technology, with Magnetic Resonance Imaging (MRI) being the gold standard due to its ability to provide superior anatomical detail, while CT scans and PET scans also play a supporting role [9], [10]. Advances in deep learning technology have brought significant changes to MRI-based brain tumor classification. However, beyond simply improving accuracy, challenges such as limited medical data, class imbalance, and the need for model interpretability remain significant issues that need to be addressed through the selection of the appropriate architecture [11], [12].

Although many architectures have shown promising results, most studies tend to focus on benchmarking to achieve the highest possible accuracy. There is a gap in research on in-depth comparative analysis of how architectural design philosophies such as sequential (VGG), residual (ResNet), and encoder-decoder (U-Net) fundamentally affect performance on the same data set. To fill this gap, this study presents a comparative analysis that aims to understand the strengths and weaknesses of each architectural design approach. We selected three models that represent three different design philosophies, with the following scientific justifications:

- VGG-16: Baseline Classical Sequential Architecture. VGG-16 was chosen because of its role as a fundamental and influential sequential architecture in deep sequential learning in the field of computer vision. Its homogeneous structure, consisting of stacks of 3x3 convolution layers, has proven to be very effective in learning hierarchical features from images [13]. Its ability to extract reliable features makes it a popular choice in transfer learning for medical tasks. In this study, VGG-16 serves as an excellent baseline for measuring the additional benefits of more complex architectures. Recent studies continue to use it as a foundation or part of hybrid models, proving its continued relevance [14].
- ResNet50: Testing the Power of Residual Connections. As network depth increases, sequential architectures such as VGG face problems of degradation and vanishing gradients. ResNet50 was chosen to directly test the hypothesis that residual connections (skip connections) can overcome these challenges, enabling the training of much deeper networks without loss of performance [15]. This capability is crucial in complex medical image analysis. Modern research continues to show that ResNet-based models excel in various medical diagnosis tasks due to their powerful and efficient feature extraction [16].
- U-Net: Evaluating Segmentation Architectures on Classification Tasks. The selection of the U-Net is central to our research hypothesis on architectural suitability. U-Net is the gold standard for biomedical image segmentation, a role in which it has demonstrated outstanding performance [17]. Its symmetric encoder-decoder architecture is explicitly designed to preserve spatial information, enabling it to produce precise pixel localization, as confirmed by numerous comprehensive reviews [18]. However, this optimization for localization may not be ideal for classification tasks, which demand abstract feature extraction. By adapting the U-Net, we can directly test the hypothesis about the importance of architectural suitability for the task.

While this study focuses on fundamental comparisons, it is important to recognize that future trends in brain tumor classification are moving towards more comprehensive systems, such as multimodal learning, explainable AI, and federated learning [19]–[21]. Based on the model selection and existing research gaps, we formulate the following hypothesis: model performance is not only determined by its complexity, but by the fundamental suitability between its architectural design and the given task. Specifically, we predict that ResNet50 will outperform VGG-16, and U-Net will show the lowest performance in this classification task.

2. Related Work

Rashid Baimukashev et al. [22] conducted research on fuzzy computer vision techniques in a biomedical context, to investigate the methodologies, applications, and challenges that exist at the intersection of fuzzy logic and computer vision, as well as guiding future research. The methodology used was a systematic survey that collected articles from the Scopus database over the past decade, focusing on fuzzy computer vision applications in biomedical image analysis, such as brain tumor detection and dental image analysis. The results identified

nine major applications, including disease detection and diagnosis, image segmentation, image classification, and image processing. They showed that the combination of fuzzy logic and deep learning can improve biomedical image analysis. Although significant progress has been made, challenges such as model interpretability and scalability remain. This research emphasizes the importance of fuzzy computer vision in improving biomedical research and proposes future research directions to overcome existing challenges and improve the application of this technique.

Adham Al-Rahbi et al. [23] compiled a systematic review of artificial intelligence (AI) applications in glioma diagnosis, assessment, genotype prediction, and treatment response, as well as challenges in AI integration into clinical practice. This study analyzed 42 relevant articles retrieved from four databases (Scopus, PubMed, Wiley, and Google Scholar) up to June 2022. The results showed that most of the articles were from the United States and China, with an increase in publications every year, especially in 2022. The main focus is on glioma, with AI applications in detection, tumor grade assessment, and survival prediction, using convolutional neural networks (CNN) as the main method. This research confirms the great potential of AI to improve the care of patients with glioma. However, challenges in its application remain, emphasizing the need for international collaboration to address such issues.

Sadegh Ghaderi et al. [24] compiled a systematic review of Diffusion Tensor Imaging (DTI) biomarkers in brain metastases (BM) and other brain tumors, such as glioblastoma. They assessed the ability of DTI and tractography in distinguishing BM from primary tumors. This study analyzed 57 studies involving 1592 BM patients and 1578 comparable tumors, with searches in PubMed, Scopus, and Web of Science databases until August 2023. Results showed that peritumoral Fractional Anisotropy (FA) consistently distinguished BM from primary tumors, while Mean Diffusivity (MD) was increased in BM. Axial and Radial Diffusivity analysis provided insight into the effects of radiation and infiltration, and tractography showed anatomical relationships important for surgical planning. This study confirms that DTI metrics can serve as non-invasive biomarkers to differentiate BM from other tumors and improve diagnosis and treatment planning, with multiparametric approaches being particularly important in brain tumor characterization.

Akkus et al. [1] discussed the use of deep learning for brain MRI segmentation, which aims to improve the accuracy of tumor detection. This research shows that CNN can produce more precise results than traditional methods, especially in detecting tumors of different shapes and sizes. In addition, challenges such as data limitations and high computational requirements were also outlined as a focus for future research.

Sunanda Das et al. [25] developed a CNN model to classify brain tumors in T1 contrast-enhanced MRI images, aiming to improve diagnosis accuracy. The dataset consists of 3064 MRI images from 233 patients, including gliomas, meningiomas, and pituitary tumors. Images were processed by resizing, smoothing using a Gaussian filter, and contrast enhancement with histogram equalization. The CNN model was trained by dividing the dataset into training, validation, and testing sets. Results show a testing accuracy of 94.39%, with an average precision of 93.33% and an average recall of 93%. The model shows good performance in identifying pituitary tumors and can potentially be used in clinical practice for early diagnosis of brain tumors.

According to Rajinikanth et al [26], the VGG architecture is well-suited for detecting objects with high-resolution spatial patterns in medical imaging. This study outlines how VGG's ability to recognize features in depth makes it a superior choice for complex image analysis, including MRI-based diagnosis.

In the research outlined by Srinivas et al [27], the VGG-16 model is utilized for brain tumor classification through a transfer learning approach using an MRI dataset. By leveraging a pre-trained architecture on the ImageNet dataset, VGG-16 demonstrates exceptional performance in detecting brain tumors, achieving an accuracy of up to 96%. To address the issue of a small dataset, data augmentation techniques such as rotation, flipping, and cropping are applied to enhance data variation and model generalization. Additionally, image intensity preprocessing and pixel normalization are employed to ensure input data consistency. The strength of this model lies in its ability to automatically extract features through the depth of the network, making it an effective tool for MRI-based brain tumor diagnosis.

Khan, H. A. et al [9] used the ResNet-50 model to detect brain tumors in a small dataset consisting of 253 MRI images. Through a transfer learning approach, ResNet-50 demonstrated an advantage in learning residual features via a deep learning architecture based on

skip connections, which helps maintain gradients as the network becomes deeper. In this experiment, ResNet-50 achieved an accuracy of 89% on the training data and 87% on the validation data, slightly below the VGG-16 model. Although its accuracy is lower compared to some other models, ResNet-50 offers advantages in computational efficiency due to its architecture, designed to prevent the vanishing gradient problem, making it suitable for applications with complex yet small datasets. This study also emphasizes the importance of fine-tuning specific layers to optimize model performance according to the characteristics of medical datasets.

The research conducted by Agrawal, P. et al. [28] focuses on the development and application of the U-Net model for brain tumor segmentation using MRI images, leveraging a convolutional network architecture designed explicitly for segmentation tasks. U-Net features a "U"-shaped structure that consists of a contracting path (encoder) to capture context by reducing the image dimensions through convolutional and pooling layers, and an expanding path (decoder) that restores the image resolution back to its original size using upsampling and convolution. Skip connections between the encoder and decoder layers allow for the preservation of local information from the earlier layers, which is crucial for achieving accurate segmentation, particularly in a medical context. This study utilized an MRI dataset from various patients to train the U-Net model to differentiate tumor tissue from healthy tissue. The evaluation results demonstrated that the model could achieve high accuracy in detecting and segmenting brain tumors, affirming the potential of U-Net as an effective tool in medical diagnosis that can assist healthcare professionals in making treatment-related decisions for tumors.

From the existing literature review, it is clear that CNN architectures such as VGG-16, ResNet50, and U-Net show significant potential in brain tumor analysis. However, most studies tend to focus on benchmarking to achieve the highest accuracy, with few conducting in-depth analyses of how differences in architectural design philosophies, such as sequential (VGG), residual (ResNet), and encoder-decoder (U-Net), fundamentally impact performance. Additionally, evaluations that specifically test segmentation architectures on pure classification tasks under the same conditions, as well as considerations of practical efficiency metrics such as computational cost, are still rare in the literature. Therefore, this study is designed to address this research gap by presenting a comparative analysis that not only evaluates accuracy but also discusses how fundamental architectural design choices impact the success of brain tumor classification tasks.

3. Methodology

Figure 1 shows the proposed experimental process to evaluate the performance of several deep learning architectures in brain tumor classification. The process starts with the collection of the dataset, which then goes through a preprocessing stage to ensure the data is in optimal condition for model training and testing. The dataset is further divided into two subsets: Train and Test, to train the model and test the generalizability of the trained model. Three deep learning architectures, namely VGG-16, ResNet50, and U-Net, were used in the training stage, with a focus on exploring the ability of each architecture to handle the brain tumor classification task. The model is evaluated using various metrics, such as accuracy, sensitivity, specificity, and Dice Score, to assess its performance.

3.1. Dataset

This research used public datasets sourced from Kaggle (https://www.kaggle.com/datasets/shreyag1103/brain-mri-scans-for-brain-tumor-classification). This dataset consists of 1,311 high-resolution brain MRI (Magnetic Resonance Imaging) scans curated specifically for the task of brain tumor detection and classification. Each MRI image was labelled into one of four different classes: "pituitary," 'meningioma,' 'glioma,' or 'no tumor,' representing a range of conditions relevant to clinical applications. The class distribution of the dataset is visualized in Figure 2, which shows the relative proportion of each tumor type in the dataset. For a sample of the dataset, see Figure 3.

3.2. Preprocessing

Before proceeding to the model training stage, the dataset containing brain MRI scans must go through a preprocessing stage. This stage is crucial for ensuring the data used is of

high quality and meets the deep learning model's application needs. The preprocessing process includes several main steps, namely image reading, resizing, normalization, labeling, and dataset division. The flow of preprocessing can be seen in Figure 4.

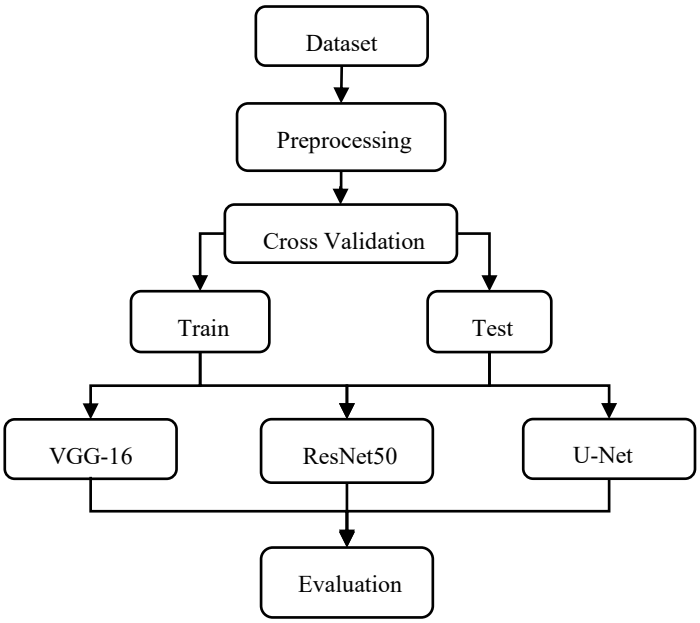


Figure 1. Experimental Methodology

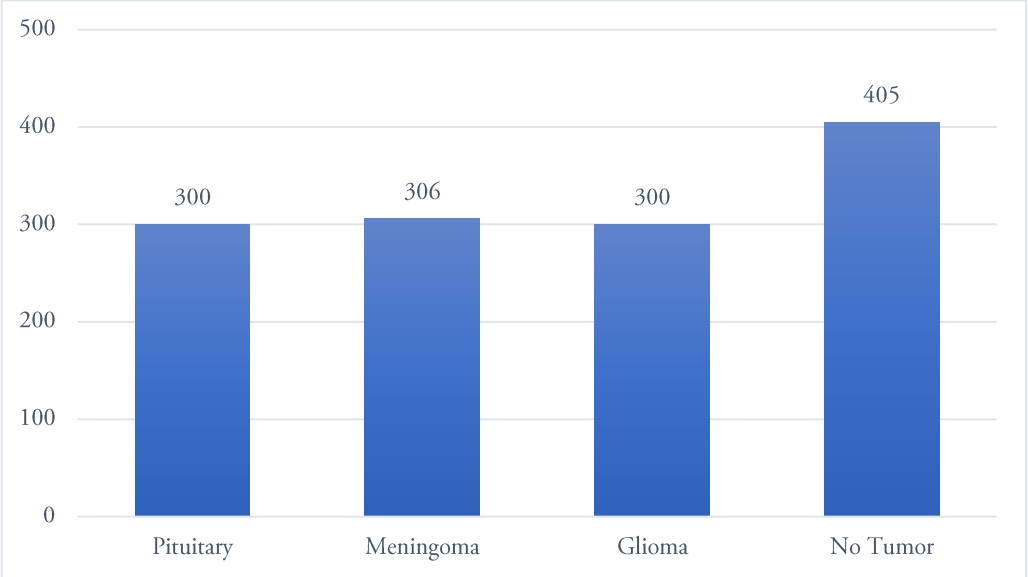


Figure 2. Distribution of Brain MRI Categories

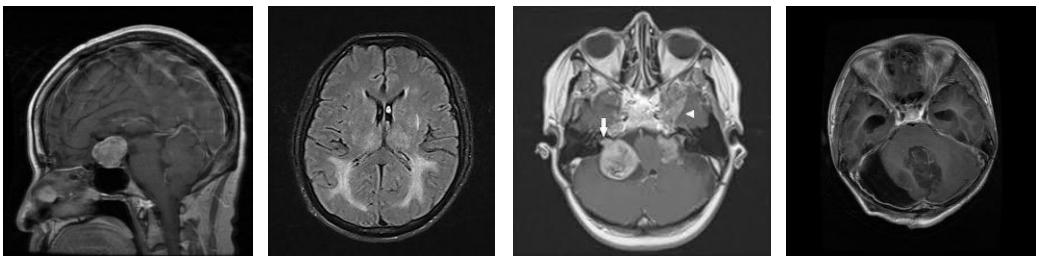


Figure 3. Sample of dataset (**a**) Pituitary; (**b**) No Tumor; (**c**) Meningioma; (**d**) Glioma.

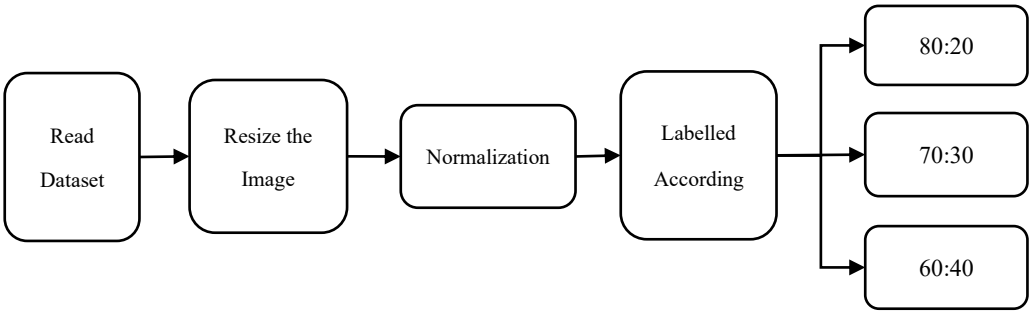


Figure 4. Preprocessing stages

As seen from Figure 4, the preprocessing stage in this research is as follows:

1. Images are read from a predefined directory. Each image is retrieved based on its file path and grouped according to its class ("glioma," "meningioma," "pituitary," "no tumor").
2. Each image is resized to a specific uniform dimension for each architecture: 224x224 pixels for VGG-16 and U-Net, and 128x128 pixels for ResNet50. This size normalization is important for processing efficiency and compatibility with each model's input layer.
3. The pixel values of each image are normalized to match the standards expected by the pre-trained model. This process involves adjusting pixel values based on statistics from the dataset, which is important for ensuring data consistency and maximizing the effectiveness of transfer learning.
4. For classification tasks, deep learning models require target variables to be in a numerical format. Consequently, the dataset's nominal labels (e.g., 'glioma', 'meningioma') are converted into a numerical format using the one-hot encoding method. This process represents each class as an orthogonal binary vector (e.g., [1, 0, 0, 0] for 'glioma'), which serves as the ground truth target for the model's loss function calculation during the training and optimization process.
5. The dataset is divided into training and testing subsets. To test the generalization of the model, the dataset was divided using three schemes. The main scheme was fold cross-validation (80:20), in which one of the five folds (20% of the data) was alternately used as test data, while the remaining four folds (80% of the data) were used as training data, as illustrated in Table 1. Additionally, to test the model with fewer training data, we also applied a single training-test split scheme with ratios of 70:30 and 60:40, whose distributions are summarized in Tables 2 and 3.

Table 1. Illustration of distribution of fold 80:20.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
20%	20%	20%	20%	20%

Table 2. Illustration of distribution of fold 70:30.

Fold 1	Fold 2	Fold 3	Rest
30%	30%	30%	10%

Table 3. Illustration of distribution of fold 60:40.

Fold 1	Fold 2	Rest
40%	40%	20%

Note: for the 70:30 scheme, if one-fold (30%) is used as test data, then the other two folds plus "Rest" (30% + 30% + 10% = 70%) are combined into training data. The same applies to the 60:40 scheme.

3.3. Comparative Models

In this research, several promising CNN architectures will be selected for implementation. These models will be selected due to their proven effectiveness in image classification and segmentation tasks, especially in the context of medical imaging. The selection process will consider various factors, including architectural complexity, ability to capture complex features, and model performance in similar studies. The selected CNN models will include established architectures such as VGG-16, ResNet50, and U-Net. A comparative study is conducted to evaluate the performance of these models in brain tumor image classification and segmentation tasks.

3.3.1. VGG-16

VGG (Visual Geometry Group) is a CNN architecture known for its simple yet efficient design. It uses convolution layers with deeply structured 3x3 filters to capture important features in images. Research shows that VGG-16 and VGG-19 are effective in medical image classification, including brain MRI, through a transfer learning approach that helps overcome data limitations. In the brain tumor classification task, VGG can recognize specific patterns in brain tissue with high accuracy [29]. In this study, the model used is VGG-16, a variant of the VGG architecture that has 16 trainable layers (convolutional and fully connected). VGG-16 is renowned for its deep feature extraction capabilities, making it ideal for brain tumor classification tasks from MRI images.

Table 4. Architecture of VGG-16.

Block	Internal Layers	Output Feature Map Size
Input	-	224 ×224×3
Block 1	(2×) 3×3 Conv, 64 filters	224 ×224×64
	2 ×2 Max Pooling	112 ×112×64
Block 2	(2×) 3×3 Conv, 128 filters	112 ×112×128
	2 ×2 Max Pooling	56 ×56×128
Block 3	(3×) 3×3 Conv, 256 filters	56 ×56×256
	2 ×2 Max Pooling	28 ×28×256
Block 4	(3×) 3×3 Conv, 512 filters	28 ×28×512
	2 ×2 Max Pooling	14 ×14×512
Block 5	(3×) 3×3 Conv, 512 filters	14×14×512
	2 ×2 Max Pooling	7 ×7×512
Classification Head	Flatten	25088
	Fully Connected (Dense)	4094
	Fully Connected (Dense)	4096
	Fully Connected (Dense), Softmax	4 (number of classes)

The ReLU (Rectified Linear Unit) activation function is commonly used in intermediate dense layers to introduce non-linearity. This ensures that the network can learn complex patterns in the data by activating only certain neurons, which effectively overcomes the problem of missing gradients in deeper networks. The Softmax activation function is applied when performing multi-class classification. Softmax normalizes the output into probability distributions across multiple classes, making it possible for the model to predict the class with the highest probability for each input. This combination of ReLU and Softmax ensures robust feature extraction and accurate classification results. See Equations (1) and (2).

$$ReLU = \max(0, i) \tag{1}$$

$$softmax(x_m) = \frac{e^{Xm}}{\sum_{n=1}^{k} e^{Xn}} \tag{2}$$

The optimization process in this research utilizes the Adaptive Moment Estimation (Adam) algorithm. Adam is a stochastic optimization method that combines the benefits of both the first and second moments of the gradient to adjust model parameters efficiently.

This approach provides adaptive learning rates for each parameter, enhancing optimization stability and convergence. The mathematical formulation of Adam is represented by Equations (3)- (7).

$$m1 = beta1 \times (1 - beta1) \times gradient \tag{3}$$

$$m2 = beta2 \times m2 + (1 - beta2) \times gradient \tag{4}$$

$$m1_{hat} = \frac{m1}{(1 - beta1^{t})} \tag{5}$$

$$m2_{hat} = \frac{m2}{(1 - beta2^{t})} \tag{6}$$

$$p = \frac{p - lr \times m1_{hat}}{(\sqrt{m2_{hat}} + epsilon)} \tag{7}$$

Where, the momentum term (m) smooths updates by incorporating past gradients, while β controls how past information influences current updates. Gradients guide parameter adjustments, with p as the parameter being optimized and t as the current iteration. The learning rate (lr) sets the update step size, balancing speed and stability. Epsilon (ϵ) prevents division by zero, ensuring numerical stability.

3.3.2. ResNet50

ResNet (Residual Network) introduces residual connections to overcome the vanishing gradient problem that often occurs in very deep networks. This approach allows the model to train a network with a larger number of layers without losing performance, while improving the ability to capture complex features from MRI images. In medical image classification tasks, ResNet has proven to excel, especially for high-complexity datasets such as brain MRI, where it can detect different types of tumors based on subtle visual feature differences [30].

Table 5. Architecture of ResNet50.

Block	Internal Layers	Output Feature Map Size	Skip Connection
Input	-	128 ×128×3	None
Block 1	7×7 Conv, 64 filters, Stride 2	64 ×64×64	None
	3 ×3 Max Pooling. Stride 2	32 ×32×64	None
Block 2 (conv2_x)	Block Bottleneck, [1x1,64; 3x3,64; 1x1,256] ×3	32 ×32×256	1 Projection 2 Identity
Block 3 (conv3_x)	Block Bottleneck, [1x1,128; 3x3,128; 1x1,512] ×4	16 ×16×512	1 Projection 3 Identity
Block 4 (conv4_x)	Block Bottleneck, [1x1,256; 3x3,256; 1x1,1024] ×6	8 ×8×1024	1 Projection 5 Identity
Block 5 (conv5_x)	Block Bottleneck, [1x1,512; 3x3,512; 1x1,2048] ×3	4×4×2048	1 Projection 2 Identity
Classification Head	Global Average Pooling	2048	None
	Fully Connected (Dense), SoftMax	4 (number of classes)	None

In this study, ResNet50 is used, a variant of the ResNet architecture that has 50 trainable layers. ResNet50 is designed to optimize classification by utilizing residual connections to accelerate training convergence and improve model accuracy. The network structure of ResNet50 allows for more in-depth feature exploration, making it particularly suitable for brain tumor classification in MRI images. ResNet50 also uses a ReLU (Rectified Linear Unit) activation function (Equation (1)) in the hidden layer to add an element of non-linearity, as well as a Softmax activation function (Equation (2)) in the output layer to support multi-class classification. In addition, the model is also optimized using Adam's algorithm, which helps

in accelerating training by combining momentum and gradient mean square, ensuring more stable and efficient parameter updates. The network structure of ResNet50 allows for more in-depth feature exploration, making it particularly suitable for brain tumor classification in MRI images.

3.3.3. U-Net

U-Net is a CNN architecture specifically designed image segmentation, with a primary focus on segmenting images with high precision. The model uses an encoder-decoder structure equipped with skip connections [31]. The encoder part is responsible for extracting features from the input image. In this section, the image is processed through successive convolutional layers followed by pooling operations to reduce the image's dimensions and capture high-level features. This step helps to understand the broader context of the image by capturing abstract representations of the data.

Table 6. Architecture of the U-Net encoder.

Level	Internal Layers	Number of Channels	Output Feature Map Size
Input	-	1	256 ×256
1	(2×) 3×3 Conv + ReLU	64	256 ×256
	2 ×2 Max Pooling	64	128 ×128
2	(2×) 3×3 Conv + ReLU	128	128 ×128
	2 ×2 Max Pooling	128	64×64
3	(2×) 3×3 Conv + ReLU	256	64×64
	2 ×2 Max Pooling	256	32×32
4	(2×) 3×3 Conv + ReLU	512	32×32
	2 ×2 Max Pooling	512	16×16
5 (Bottleneck)	(2×) 3×3 Conv + ReLU	128	16×16

The decoder part is responsible for reconstructing the image to its original resolution, using up-sampling layers that increase the image's dimensions and recover details lost during down-sampling in the encoder. Skip connections link corresponding layers in the encoder and decoder to transfer detailed information that might be lost in the pooling stages. This feature is especially useful in segmentation tasks, as it ensures that spatial details, critical for accurate segmentation, are retained throughout the process.

Table 7. Architecture of the U-Net decoder.

Level	Internal Layers	Number of Channels	Output Feature Map Size
4	2×2 Up-Conv	512	32×32
	Concatenate with output from Encoder Level 4	1024	32×32
	(2×) 3×3 Conv + ReLU	512	32×32
3	2×2 Up-Conv	256	64×64
	Concatenate with output from Encoder Level 3	512	64×64
	(2×) 3×3 Conv + ReLU	256	64×64
2	2×2 Up-Conv	128	128×128
	Concatenate with output from Encoder Level 2	256	128×128
	(2×) 3×3 Conv + ReLU	128	128×128
1	2×2 Up-Conv	64	256 ×256
	Concatenate with output from Encoder Level 1	128	256 ×256
	(2×) 3×3 Conv + ReLU	64	256 ×256

3.4. Evaluation

Evaluation metrics are essential for assessing the classification performance of a trained model. In this study, the metrics employed include accuracy, precision, recall (or sensitivity),

and F1-Score. Accuracy represents the proportion of correctly classified instances out of the total evaluated dataset (8). Precision measures the ratio of true positive predictions to the total positive predictions made by the model (9). Recall, or sensitivity, evaluates the model's ability to correctly identify positive instances out of all actual positive cases (10). The F1-Score combines precision and recall into a single metric, providing a harmonic mean to balance both measures, particularly useful when dealing with imbalanced datasets (11). These metrics are calculated using the values of TP (true positives), FP (false positives), TN (true negatives), and FN (false negatives).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precission = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1\ Score = \frac{2 \times Precission \times Recall}{Precission + Recall} \tag{11}$$

4. Experimental Results and Discussion

The implementation was carried out on the Kaggle platform, utilizing a P-100 GPU with 16 GB of memory and 30 GB of RAM. The experiments were conducted in Python, leveraging libraries such as TensorFlow, Keras, NumPy, scikit-learn, Matplotlib, and Pandas. The experimental process consisted of three stages:

- The first stage involved 80:20 cross-validation, with training and testing performed for each fold.
- The second stage used 70:30 cross-validation, again with training and testing for each fold.
- The third stage returned to 80:20 cross-validation, repeating the training and testing process for each fold.

4.1. Evaluation of 80:20 Cross-Validation

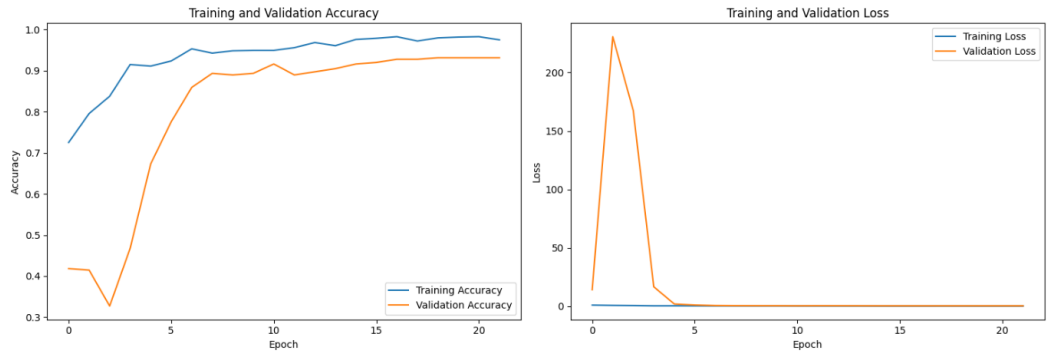
Tests with an 80:20 ratio were conducted to evaluate the models' performance using a dominant proportion of training data, focusing on accuracy, precision, recall, and F1-Score. ResNet50 consistently outperformed other models across all metrics, achieving an average accuracy of 96.42%, precision of 97.04%, recall of 96.95%, and F1-Score of 96.93%, with values ranging from 91.95% to 99.26% across folds. VGG-16 also demonstrated strong performance, with average accuracy, precision, recall, and F1-Score values of 96.34%, 93.90%, 93.75%, and 93.90%, respectively, showing only a small difference compared to ResNet50. In contrast, U-Net exhibited significantly lower performance, with an average accuracy of 82.61%, precision of 82.78%, recall of 81.69%, and F1-Score of 79.86%, consistently underperforming across all folds. For details of each evaluation, see Table 8 and Figure 8.

Table 8. Evaluation of Cross-Validation 80:20

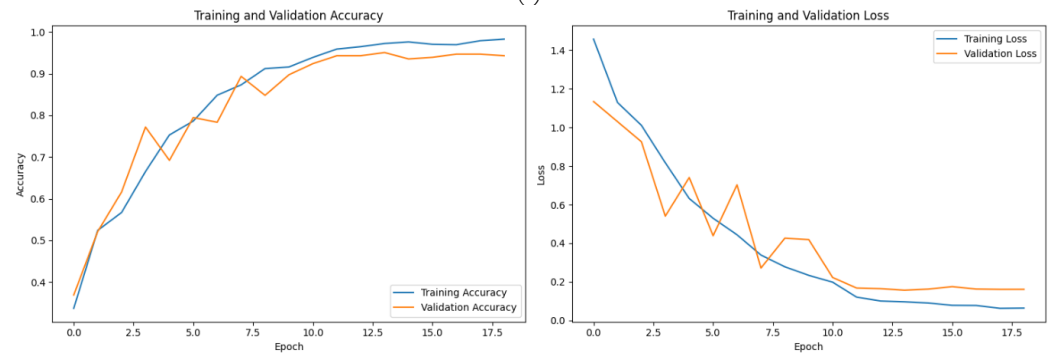
Model	Fold	Accuracy	Precision	Recall	F1-Score
U-Net	1	82.89	84.84	84.41	83.07
	2	79.77	76.91	74.61	72.66
	3	82.44	78.36	77.86	73.98
	4	83.97	86.04	85.11	84.23
	5	83.97	87.74	86.26	85.38
Means		82.61	82.78	81.69	79.86
VGG-16	1	90.87	88.52	88.21	87.93
	2	95.42	90.86	90.84	90.55
	3	98.47	93.75	93.51	93.45

Table 8 (cont.)

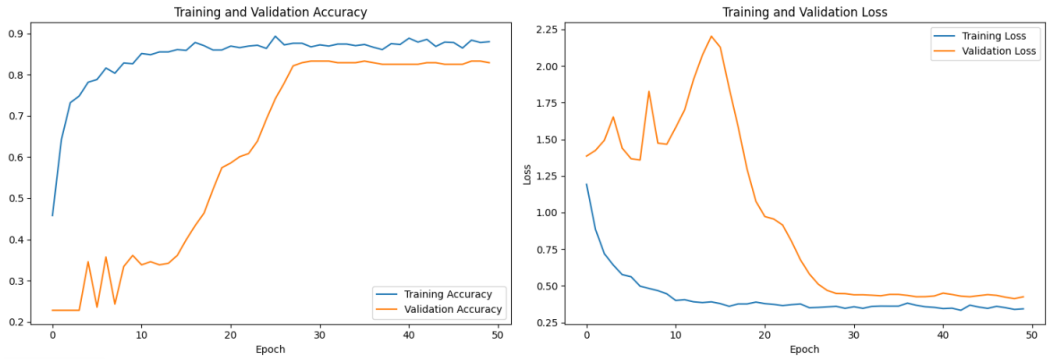
Model	Fold	Accuracy	Precision	Recall	F1-Score
	4	98.47	98.89	98.85	98.85
	5	98.47	97.49	97.33	97.31
Means		96.34	93.90	93.75	93.90
ResNet50	1	90.87	92.23	92.02	91.95
	2	98.09	99.26	99.24	99.24
	3	97.33	97.71	97.71	97.70
	4	97.71	97.08	96.95	96.90
	5	98.09	98.91	98.85	98.85
Means		96.42	97.04	96.95	96.93



(a)



(b)



(c)

Figure 8. Sample of epoch plots 80:20; (**a**) ResNet50; (**b**) VGG-16; (**c**) U-Net.

4.2. Evaluation of 70:30 Cross-Validation

In the 70:30 test, where model performance was evaluated with more test data, a decline in average accuracy was observed compared to the 80:20 test, particularly for U-Net, which achieved an average accuracy of 81.85%. VGG-16 and ResNet50 maintained high performance with average accuracies of 94.89% and 94.97%, respectively, demonstrating their adaptability to varying data proportions. Precision testing showed a similar trend, with U-Net

dropping to an average precision of 80.87%, while VGG-16 and ResNet50 achieved 94.02% and 94.44%, respectively. Recall results further confirmed the robustness of VGG-16 and ResNet50, each with an average recall of 93.82%, whereas U-Net showed a decline to 80.64%. Lastly, F1-Score testing revealed a significant gap, with U-Net scoring an average of 79.45%, while VGG-16 and ResNet50 maintained strong performance at 93.30% and 93.72%, respectively. These results highlight the superior adaptability of VGG-16 and ResNet50 to changes in data proportion, while U-Net's performance is notably more impacted. For details of each evaluation, see Table 9 and Figure 9.

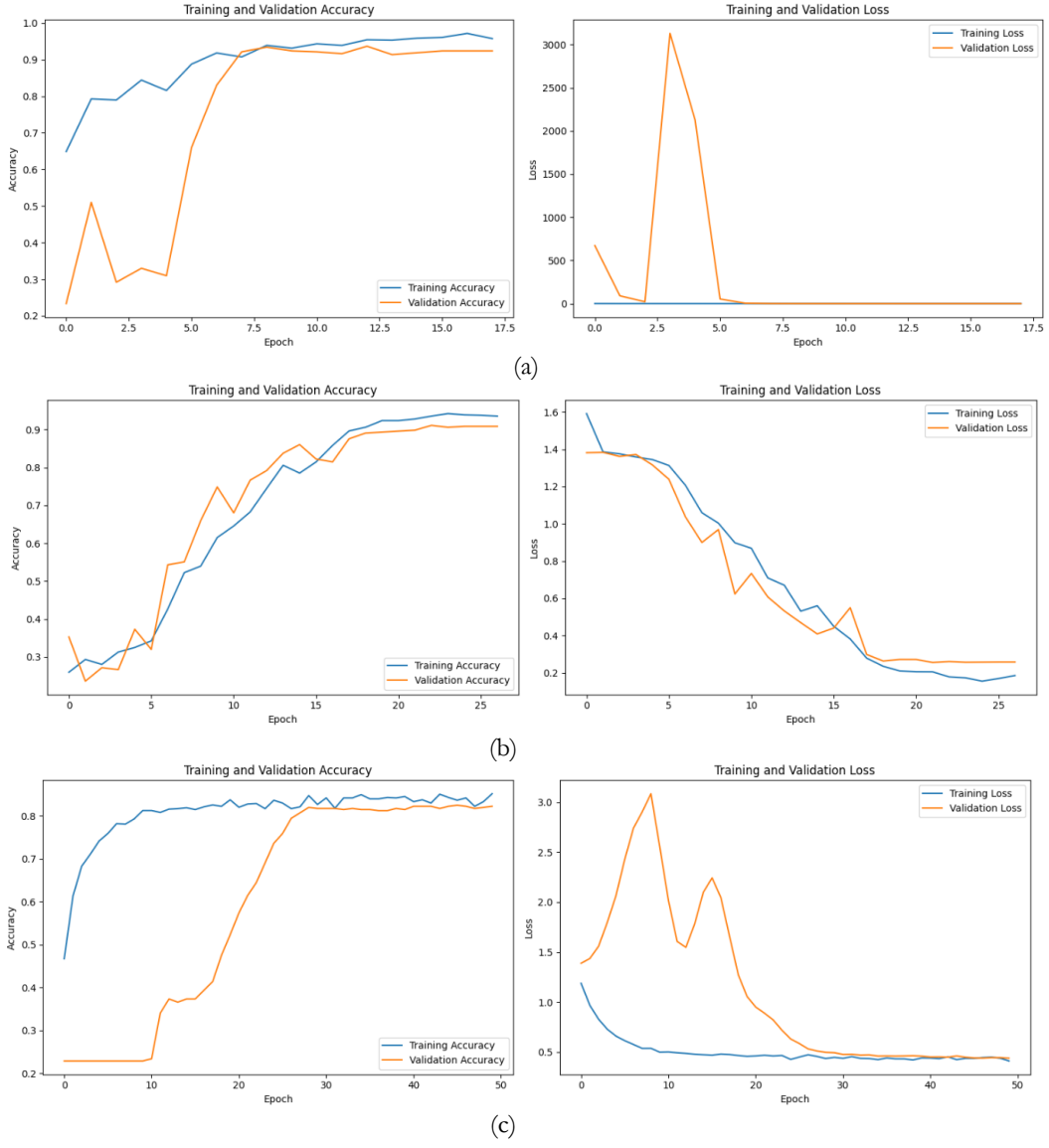


Figure 9. Sample of epoch plots 70:30; (**a**) ResNet50; (**b**) VGG-16; (**c**) U-Net.

Table 9. Evaluation of Cross-Validation 70:30

Model	Fold	Accuracy	Precision	Recall	F1-Score
U-Net	1	81.01	78.46	79.91	78.06
	2	80.78	82.38	81.46	79.96
	3	83.75	81.78	80.55	80.35
Means		82.61	81.85	80.87	80.64
VGG-16	1	91.76	92.18	91.99	91.90
	2	96.34	94.75	94.74	94.72
	3	96.57	95.14	94.74	94.70

Table 9 (cont.)

Model	Fold	Accuracy	Precision	Recall	F1-Score
Means		96.34	94.89	94.02	93.82
ResNet50	1	91.76	91.14	90.16	89.94
	2	97.71	96.56	96.34	96.31
	3	95.42	95.67	94.97	94.91
Means		96.42	94.97	94.44	93.82

4.3. Evaluation of 60:40 Cross-Validation

The final test, using a 60:40 ratio with the highest proportion of test data, provided significant challenges for the model. The U-Net model showed a notable drop in performance, with an average accuracy of 72.57%, precision of 75.55%, recall of 72.00%, and F1-Score of 69.61%, highlighting its difficulty in adapting to this scenario. In contrast, the VGG-16 model maintained strong results despite the downturn, achieving an average accuracy of 90.57%, precision of 93.56%, recall of 93.43%, and F1-Score of 93.40%, demonstrating its reliability. ResNet50 outperformed both models, with an average accuracy of 94.67%, precision of 95.05%, recall of 95.04%, and F1-Score of 94.96%, reaffirming its superior robustness and adaptability when handling a larger proportion of test data. For details of each evaluation, please refer to Table 10 and Figure 10.

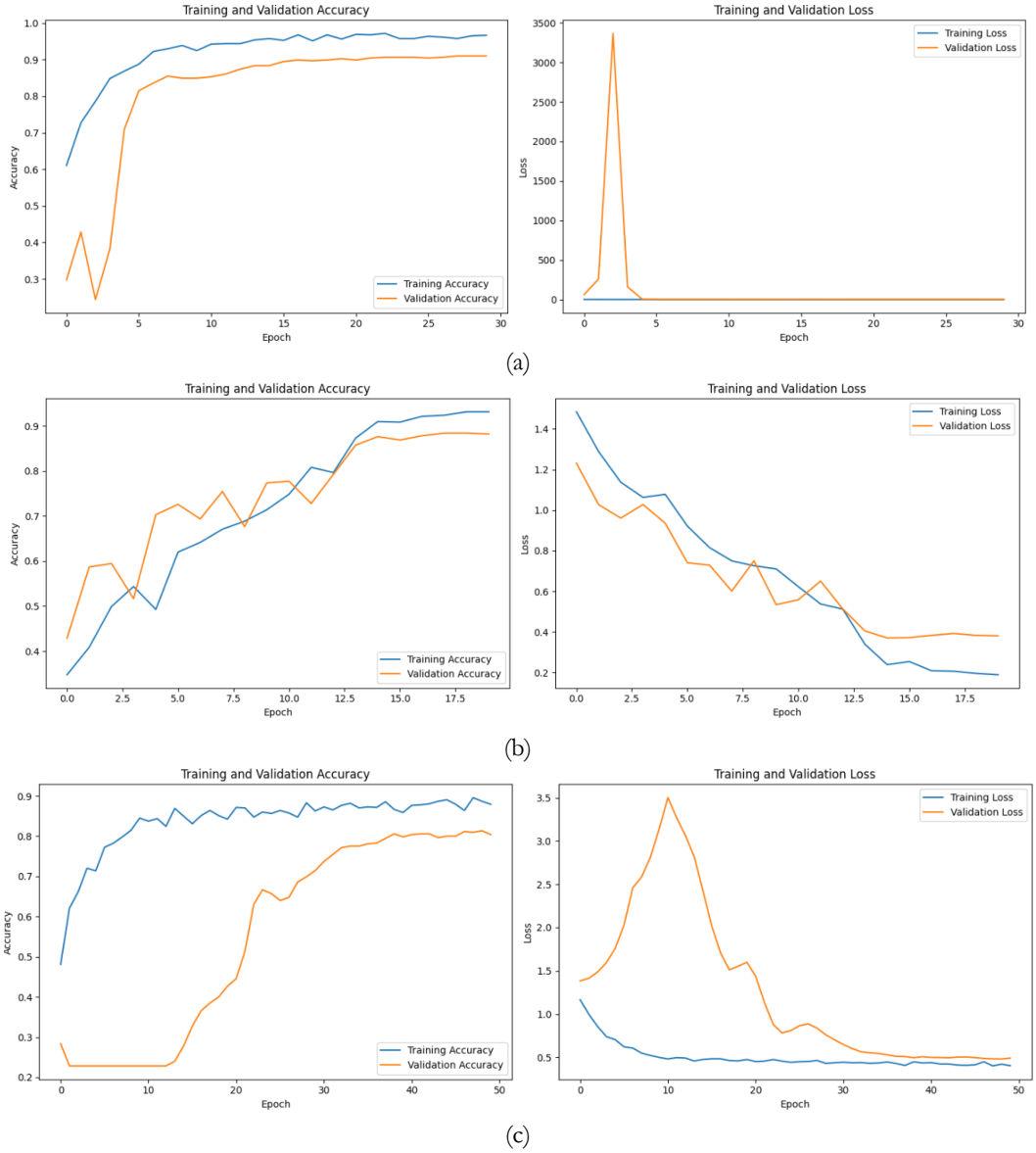


Figure 10. Sample of epoch plots 60:40; (**a**) ResNet50; (**b**) VGG-16; (**c**) U-Net.

Table 10. Evaluation of Cross-Validation 60:40

Model	Fold	Accuracy	Precision	Recall	F1-Score
U-Net	1	74.10	74.90	71.81	69.44
	2	71.05	76.21	72.19	69.78
Means		82.61	72.57	75.55	72.00
VGG-16	1	87.62	92.25	92.19	92.15
	2	93.52	94.87	94.67	94.65
Means		96.34	90.57	93.56	93.43
ResNet50	1	91.81	92.17	92.19	92.05
	2	97.52	97.94	97.90	97.88
Means		96.42	94.67	95.05	95.04

4.4. Computational Cost and Efficiency Analysis

Beyond classification performance, the practical feasibility of a model for clinical applications also heavily depends on its computational efficiency. To analyze the trade-off between accuracy and resource costs, we evaluated each architecture based on several key metrics: training time, number of parameters, and model size. A comprehensive comparison of these efficiency metrics for each testing scenario is presented in Table 11.

Table 11. Evaluation of Cross-Validation 70:30

Split Ratio	Model	Number of Parameters (millions)	Model Size (MB)	Training Time (minutes)
80:20	U-Net	7.76	89.04	10.11
	VGG-16	14.98	171.52	3.16
	ResNet50	26.71	306.14	2.25
70:30	U-Net	7.76	89.04	9.47
	VGG-16	14.98	171.52	3.20
	ResNet50	26.71	306.14	2.16
60:40	U-Net	7.76	89.04	8.36
	VGG-16	14.98	171.52	2.09
	ResNet50	26.71	306.14	2.48

4.5. Overall Evaluation

This section, summarizes the average of all evaluations performed on each model based on tests conducted with various data shares (80:20, 70:30, and 60:40). The average performance metrics, including accuracy, precision, recall, and F1-score, are calculated from the average values in each data sharing scenario to provide an overview of each model's performance, see Table 12.

Table 12. Average of the evaluation.

Model	Fold	Accuracy	Precision	Recall	F1-Score
U-Net	80:20	82.61	82.78	81.69	79.86
	70:30	81.85	80.87	80.64	79.45
	60:40	72.57	75.55	72.00	69.61
Means		79.00	79.73	78.11	76.31
VGG-16	80:20	96.34	93.90	93.75	93.90
	70:30	94.89	94.02	93.82	93.30
	60:40	90.57	93.56	93.43	93.40
Means		93,93	93,82	93,66	93,43
ResNet50	80:20	96.42	97.04	96.95	96.93
	70:30	94.97	94.44	93.82	93.72

Table 12 (cont.)

Model	Fold	Accuracy	Precision	Recall	F1-Score
	60:40	94.67	95.05	95.04	94.96
Means		95,35	95,51	95,27	95,20

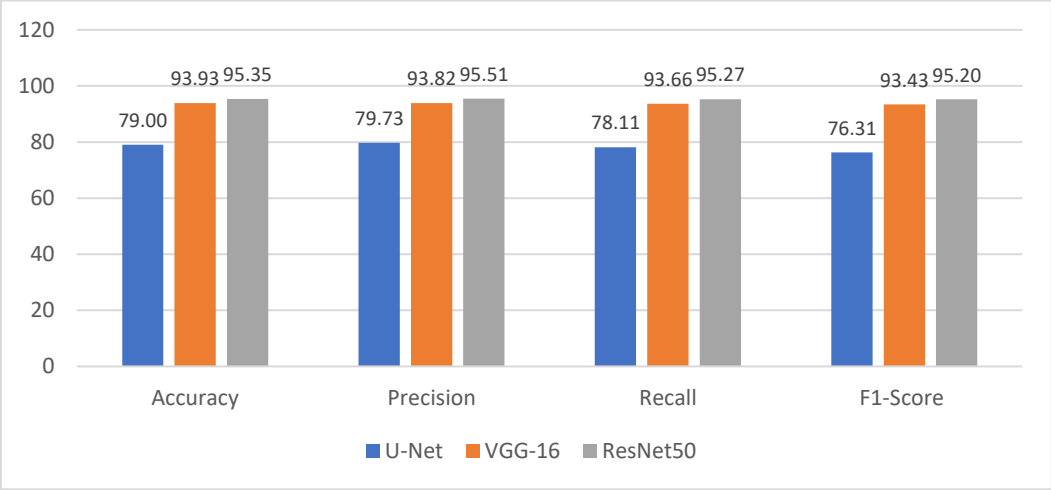


Figure 11. Chart the average of the evaluation

The results show that ResNet50 is the highest performing model across all metrics. Specifically, ResNet50 achieved an average accuracy of 95.35%, precision of 95.51%, recall of 95.27%, and F1-score of 95.2%. These results highlight the advantages of the ResNet50 architecture, which addresses deep network challenges through the use of skip connections that prevent the loss of important information.

On the other hand, VGG-16 also showed strong performance, achieving an average accuracy of 93.93%, precision of 93.82%, recall of 93.66%, and F1-score of 93.43%. The performance of VGG-16is close to that of ResNet50, demonstrating its effectiveness in the brain tumor classification task despite lacking the jump connections that enhance ResNet50 capabilities.

Meanwhile, U-Net lags behind the other two models, with an average accuracy of 79%, precision of 79.73%, recall of 78.11%, and F1 score of 76.31%. Although U-Net has proven to be effective in segmentation tasks, its performance in brain tumor classification is much lower compared to VGG-16 and ResNet50.

Overall, ResNet50 outperforms VGG-16 and U-Net in all metrics, demonstrating its robustness and efficiency in handling brain tumor classification tasks. VGG-16 follows closely, while U-Net's performance highlights the need for further optimization in classification-specific tasks.

4.6. Discussion

The evaluation results confirm our initial hypothesis that the alignment between architectural design and the given task is more crucial than model complexity alone. ResNet50 consistently achieved the highest accuracy across all scenarios, largely due to its skip connection mechanism, which prevents information degradation in deep networks and enhances the model's ability to capture essential features without losing detail.

Although VGG-16 also performed well, its lack of skip connections makes it fully dependent on sequential propagation, which may cause performance degradation in deeper architectures. However, its simpler structure allows for significantly faster training compared to ResNet50, making it more efficient in terms of computational cost.

On the other hand, U-Net shows the weakest performance. Its encoder–decoder design reduces image resolution through pooling, leading to potential information loss. While the decoder attempts to recover this information through up-convolution, some critical details cannot be fully restored, and new irrelevant data may be introduced. This explains why U-Net is less effective for brain tumor classification than ResNet50 and VGG-16.

The performance trend analysis in Table 8 shows a very interesting pattern. In scenarios with abundant training data (80:20 and 70:30), VGG-16 and ResNet50 show very competitive

and relatively equal performance. However, ResNet50's significant advantage appears in the 60:40 scenario, where the proportion of training data is the least. The accuracy of VGG-16 drops dramatically to 90.57%, while ResNet50 is able to maintain high performance at 94.67%. This indicates that the sequential architecture of VGG-16 is more susceptible to performance degradation when faced with data limitations. In contrast, ResNet50 shows superior robustness in conditions with less data.

The superior stability of ResNet50 can be attributed to its residual connection (skip connection) mechanism. These connections not only overcome the vanishing gradient problem but also allow the model to learn more efficiently from limited data by maintaining information flow. The practical implications of this finding are significant for clinical applications. In the real world, the availability of labeled MRI data is often limited. A robust model that maintains high performance even with fewer data points, such as ResNet50, is far more valuable and reliable for clinical implementation.

When compared to related work, such as Llaca-Sánchez et al. [32] who reported 89.17% accuracy with a custom CNN, our transfer learning approaches achieved substantially higher results: 95.35% for ResNet50 and 93.93% for VGG-16. This highlights the advantage of transfer learning over training models from scratch.

Beyond accuracy, our computational cost analysis (Table 7) reveals an important trade-off. While ResNet50 offers the highest accuracy, VGG-16 shows a favorable balance with faster training. These findings suggest that model selection for clinical implementation should not rely solely on accuracy, but on a balanced consideration of accuracy, training efficiency, and resource availability. ResNet50 is the most suitable choice when the highest accuracy is essential, while VGG-16 may be preferred in contexts where faster computation is a priority.

5. Conclusions

This study successfully achieved its goal of conducting an in-depth comparative analysis of deep learning architectural design philosophies, which consistently showed that ResNet50 delivered superior performance. These findings strongly support our hypothesis that architectural suitability for specific tasks is a more critical success factor than model complexity alone, while also confirming that the residual connections in ResNet50 make it the most reliable and robust choice for this medical image classification task. However, this study has several limitations, particularly the evaluation being conducted on only one public dataset, which may not fully represent the diversity of real-world clinical data. Now that this comparative study has identified ResNet50 as the most effective base model, future research can be directed toward maximizing its potential. The next logical step is to conduct an in-depth exploration of the ResNet50 architecture itself, with the aim of further improving accuracy, precision, and other performance metrics. These optimization efforts may include systematic hyperparameter tuning, testing various optimizer algorithms and learning rate schedulers, and applying more advanced data augmentation techniques to improve model generalization. By focusing on refining the best models, further research may yield more mature solutions that are ready for testing in clinical scenarios.

Author Contributions: Conceptualization, M.N.E.F. and A.K.M.; Methodology, M.N.E.F. and A.K.M; Software, M.N.E.F.; Validation, A.K.M. S.W. and H.B.; Formal analysis, M.N.E.F. and A.K.M; Investigation, M.N.E.F.; Resources, M.N.E.F.; Data curation, M.N.E.F.; Writing—original draft preparation, M.N.E.F.; Writing—review and editing, A.K.M., S.W., and H.B.; Visualization, M.N.E.F.; Supervision, A.K.M., S.W., and H.B.; Project administration, A.K.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The dataset analyzed during the current study is publicly available in the Kaggle repository at https://www.kaggle.com/datasets/shreyag1103/brain-mri-scans-for-brain-tumor-classification.

Acknowledgments: The first author is deeply grateful to Dian Nuswantoro University (UDINUS) and Universiti Teknikal Malaysia Melaka (UTeM) for the invaluable opportunity to conduct this research at UTeM as part of a student mobility program.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions," *J. Digit. Imaging*, vol. 30, no. 4, pp. 449–459, Aug. 2017, doi: 10.1007/s10278-017-9983-4.

[2] D. N. Louis *et al.*, "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary," *Acta Neuropathol.*, vol. 131, no. 6, pp. 803–820, Jun. 2016, doi: 10.1007/s00401-016-1545-1.

[3] A. Işın, C. Direkoğlu, and M. Şah, "Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods," *Procedia Comput. Sci.*, vol. 102, pp. 317–324, 2016, doi: 10.1016/j.procs.2016.09.407.

[4] G. S. Tandel *et al.*, "A Review on a Deep Learning Perspective in Brain Cancer Classification," *Cancers (Basel).*, vol. 11, no. 1, p. 111, Jan. 2019, doi: 10.3390/cancers11010111.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *arXiv*, Apr. 2015, pp. 1–14. doi: 10.48550/ARXIV.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[7] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3–4, p. 100004, Sep. 2019, doi: 10.1016/j.array.2019.100004.

[8] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Z. Med. Phys.*, vol. 29, no. 2, pp. 102–127, May 2019, doi: 10.1016/j.zemedi.2018.11.002.

[9] H. Ali Khan, W. Jue, M. Mushtaq, and M. Umer Mushtaq, "Brain tumor classification in MRI image using convolutional neural network," *Math. Biosci. Eng.*, vol. 17, no. 5, pp. 6203–6216, 2020, doi: 10.3934/mbe.2020328.

[10] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 3–11. doi: 10.1007/978-3-030-00889-5_1.

[11] P. Wesseling and D. Capper, "WHO 2016 Classification of gliomas," *Neuropathol. Appl. Neurobiol.*, vol. 44, no. 2, pp. 139–150, Feb. 2018, doi: 10.1111/nan.12432.

[12] M. F. Ahamed *et al.*, "A review on brain tumor segmentation based on deep learning methods with federated learning techniques," *Comput. Med. Imaging Graph.*, vol. 110, p. 102313, Dec. 2023, doi: 10.1016/j.compmedimag.2023.102313.

[13] D. N. Louis *et al.*, "The 2021 WHO Classification of Tumors of the Central Nervous System: a summary," *Neuro. Oncol.*, vol. 23, no. 8, pp. 1231–1251, Aug. 2021, doi: 10.1093/neuonc/noab106.

[14] E. M. Younis, I. A. Ibrahim, M. N. Mahmoud, and A. M. Albarrak, "Hybrid of VGG-16 and FTVT-b16 Models to Enhance Brain Tumors Classification Using MRI Images," *Diagnostics*, vol. 15, no. 16, p. 2014, Aug. 2025, doi: 10.3390/diagnostics15162014.

[15] Q. T. Ostrom, N. Patil, G. Cioffi, K. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017," *Neuro. Oncol.*, vol. 22, no. Supplement_1, pp. iv1–iv96, Oct. 2020, doi: 10.1093/neuonc/noaa200.

[16] A. Sanapala, A. Sanapala, and P. Naveena, "A Review On Image Based Diagnosis UsingResNet-50," *Int. J. Res. Anal. Rev.*, vol. 11, no. 1, pp. 796–808, 2024, [Online]. Available: https://www.ijrar.org/papers/IJRAR24A3421.pdf

[17] P. Sabeghi *et al.*, "Advances in Neuro-Oncological Imaging: An Update on Diagnostic Approach to Brain Tumors," *Cancers (Basel).*, vol. 16, no. 3, p. 576, Jan. 2024, doi: 10.3390/cancers16030576.

[18] W. Jiangtao, N. I. R. Ruhaiyem, and F. Panpan, "A Comprehensive Review of U-Net and Its Variants: Advances and Applications in Medical Image Segmentation," *IET Image Process.*, vol. 19, no. 1, Jan. 2025, doi: 10.1049/ipr2.70019.

[19] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[20] D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 221–248, Jun. 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[21] R. Kaifi, "A Review of Recent Advances in Brain Tumor Diagnosis Based on AI-Based Classification," *Diagnostics*, vol. 13, no. 18, p. 3007, Sep. 2023, doi: 10.3390/diagnostics13183007.

[22] R. Baimukashev, S. Kadyrov, and C. Turan, "Systematic Survey of Deep Fuzzy Computer Vision in Biomedical Research," *Fuzzy Inf. Eng.*, vol. 16, no. 3, pp. 220–243, Sep. 2024, doi: 10.26599/FIE.2024.9270043.

[23] A. Al-Rahbi, O. Al-Mahrouqi, and T. Al-Saadi, "Uses of artificial intelligence in glioma: A systematic review," *Med. Int.*, vol. 4, no. 4, p. 40, May 2024, doi: 10.3892/mi.2024.164.

[24] U. A. Siddiqui *et al.*, "Quality assessment of critical and non-critical domains of systematic reviews on artificial intelligence in gliomas using AMSTAR II: A systematic review," *J. Clin. Neurosci.*, vol. 131, p. 110926, Jan. 2025, doi: 10.1016/j.jocn.2024.110926.

[25] S. Das, O. F. M. R. R. Aranya, and N. N. Labiba, "Brain Tumor Classification Using Convolutional Neural Network," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, May 2019, pp. 1–5. doi: 10.1109/ICASERT.2019.8934603.

[26] A. Younis, L. Qiang, C. O. Nyatega, M. J. Adamu, and H. B. Kawuwa, "Brain Tumor Analysis Using Deep Learning and VGG-16 Ensembling Learning Approaches," *Appl. Sci.*, vol. 12, no. 14, p. 7282, Jul. 2022, doi: 10.3390/app12147282.

[27] C. Srinivas *et al.*, "Deep Transfer Learning Approaches in Performance Analysis of Brain Tumor Classification Using MRI Images," *J. Healthc. Eng.*, vol. 2022, pp. 1–17, Mar. 2022, doi: 10.1155/2022/3264367.

[28] P. Agrawal, N. Katal, and N. Hooda, "Segmentation and classification of brain tumor using 3D-UNet deep neural networks," *Int. J. Cogn. Comput. Eng.*, vol. 3, pp. 199–210, Jun. 2022, doi: 10.1016/j.ijcce.2022.11.001.
[29] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images," *Int. J. Sci. Res. Publ.*, vol. 9, no. 10, p. p9420, Oct. 2019, doi: 10.29322/IJSRP.9.10.2019.p9420.
[30] S. Targ, D. Almeida, and K. Lyman, "Resnet in Resnet: Generalizing Residual Architectures," *arXiv*. Mar. 25, 2016. [Online]. Available: http://arxiv.org/abs/1603.08029
[31] W. Weng and X. Zhu, "INet: Convolutional Networks for Biomedical Image Segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, 2021, doi: 10.1109/ACCESS.2021.3053408.
[32] B. A. Llaca-Sanchez, D. S. Marın-Chavez, and A. Vilchis-yubi, "Detection and Classification of Brain Tumors in MRIs Using CNNs," *Res. Comput. Sci.*, vol. 153, no. 12, pp. 5–18, 2024, [Online]. Available: https://www.rcs.cic.ipn.mx/2024_153_12/Detection and Classification of Brain Tumors in MRIs Using CNNs.pdf