Journal of Computer Science

Research Article

Cascaded MNetV3UNet: A Lightweight Two-Stage Architecture for High-Precision Brain Tumor Segmentation in MRI

Mayuri Popat[1] and Sanskruti Patel[2]

[1]*U & P. U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology, Faculty of Technology and Engineering, Charotar University of Science and Technology (CHARUSAT), Changa, India*
[2]*Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPIC), Charotar University of Science and Technology (CHARUSAT), Changa, India*

Article history
Received: 07-07-2025
Revised: 10-09-2025
Accepted: 26-09-2025

Corresponding Author:
Mayuri Popat
U & P. U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology, Faculty of Technology and Engineering, Charotar University of Science and Technology [CHARUSAT], Changa, India
Email: mayuripopat.ce@charusat.ac.in

Abstract: Precise brain tumor segmentation is an essential but quite challenging process in MRI images because of their irregular shapes, heterogeneous appearance, and low contrast with surrounding tissues. While U-Net-based architectures have achieved significant success, their high computational complexity limits deployment on resource-constrained systems. In this study, a Novel two-stage cascaded architecture, MNetV3UNet, is introduced, which employs the light-weight MobileNetV3-Large as an encoder and the standard U-Net as the decoder. MobileNetV3 block consists of a sequence of blocks that generate feature maps enhanced using Inverted Residual Blocks and squeeze-and-excitation modules. The Unet decoder consists of an iterative process of upsampling, interpolation, concatenation, and refinement. The first stage produces a coarse segmentation, which is refined in the second stage to enhance boundary accuracy and detail. The cascaded approach leverages a multi-scale feature extraction process, which is also coupled with the progressive refinement method. This way, we ensure a much higher degree of segmentation precision while still maintaining computational efficiency. The proposed model has been tested on the BraTS 2020 dataset, yielding a Dice score of 88.35 for Whole Tumor (WT), 89.03 for Tumor Core (TC), and 92.30% for Enhancing Tumor (ET). Additionally, it achieved a Jaccard score of 83.76 for WT, 86.15% for TC, and 89.68% for ET. The specificity obtained for WT was 99.72, for TC, 99.89, and for ET, 97.96%. It achieved sensitivity of 88.91 for WT, 89.37 for TC, and 92.88% for ET. These outcomes provide clear evidence of the proposed innovative architecture's ability to achieve an excellent balance between segmentation accuracy and computational efficiency.

Keywords: Brain Tumor, Deep Learning, Magnetic Resonance Imaging, Mobile NetV3 Large, Segmentation, UNet

Introduction

Any abnormal growth in the Human Skull can have major negative effects on the body as it is a very sensitive structure (Zafar et al., 2024). Within the Central Nervous System (CNS), two of the most essential parts of the body are the brain and the spinal cord (Zafar et al., 2024). These two parts participate in intricate processes that are still not entirely understood (Zafar et al., 2024). To determine which parts of the brain are affected by a tumor, segmenting a brain tumor is an essential step while processing a medical image (Ullah et al., 2023). Precise brain tumor segmentation, monitoring disease progression, therapeutic planning, and accurate diagnosis are critical tasks. It is challenging and time-consuming to manually detect brain tumors due to their complex nature and the diversity of patients (Ullah et al., 2023).

A commonly used neuroimaging technique for the quantitative evaluation of brain tumours in clinical settings is Magnetic Resonance Imaging (MRI). This method's high contrast of soft tissues, non-invasive nature, and absence of electrical radiation make it a beneficial option. MRI encompasses several imaging modalities, including Fluid Attenuation Inversion

Recovery (FLAIR), T2-weighted (T2), T1-weighted (T1), and T1-weighted with contrast enhancement (T1ce) (Yue et al., 2023). Each of these modalities provides crucial information, enabling physicians to make the most accurate diagnoses (Li et al., 2023). The outcomes produced by deep learning are astonishing when it comes to challenges involving medical image processing, including the segmentation of brain tumors.

Brain tumor segmentation has demonstrated cutting-edge performance with recent designs such as the U-Net and its variations. CNN, which is capable of learning and extracting features from images, has gained popularity among researchers recently and has demonstrated exceptional performance with an extremely high level of accuracy in segmenting images. Numerous researchers have used CNNs in MRI images to segment brain tumors automatically (Daimary et al., 2020). The U-Net model's basic structure has two primary paths and is based on a standard CNN. It is similar to an auto-encoder design, with the contracting or down sampling path on the left and the expanding or up sampling path on the right (the deconvolution and convolutional paths). To recover the original image resolution lost during the contracting path, where input images were down sampled, several approaches are employed by the expanding path, which is optimal, such as concatenating skip connections. By producing dense predictions at a higher resolution, the network gains knowledge about spatial classification along the expansion path (Yousef et al., 2023). However, U-Net's architecture may be computationally intensive, making it challenging to deploy on low-resource devices. Researchers have explored alternative lightweight architectures, including encoder-decoders, to mitigate the computational complexity of the U-Net, such as the MobileNetV3 large model. In comparison to many other CNN designs of comparable size, MobileNetV3 is lightweight and offers efficient feature extraction but may lack the accuracy required for precise medical segmentation (Alsenan et al., 2022; Amin et al., 2023).

To address this trade-off, we propose a novel two-stage cascaded framework, MNetV3UNet, that integrates the efficiency of MobileNetV3-Large and the hierarchical decoding of U-Net. The first stage provides an initial segmentation, capturing coarse tumor boundaries. The second stage refines these predictions, enabling better delineation of tumor sub-regions and boundaries. This coarse-to-fine strategy ensures accurate segmentation with reduced model complexity.

The Key contributions of this paper are as follows:

- To accurately segment brain tumors, a two-stage cascaded framework, MNetV3Unet, has been proposed, combining MobileNetV3–Large as the encoder and UNet as the decoder. The lightweight MobileNetV3-Large is used as a backbone, which can significantly reduce the computational complexity
- For better feature representation, hard-swish activation functions (h-swish) and Squeeze-and-Excitation (SE) blocks are incorporated in the encoder to enhance non-linearity and channel-wise attention
- Bilinear Interpolation is used in Skip connections at both stages before the concatenation of the down sampling layer and the corresponding up sampling layer
- Performed rigorous preprocessing, including Z-score Normalization, cropping, unwanted slice removal, and multi-modal input stacking (T1, T1ce, T2, FLAIR), ensuring high-quality and consistent data input. Additionally, data augmentation techniques such as vertical and horizontal flipping are applied to improve model generalization against various Brain MRI Images.
- Conducted a detailed comparison of various optimizers, including Adam, AdamW, RMSProp, and SGD, to study their impact on segmentation performance. The model's effectiveness was validated on the BRATS 2020 dataset using standard evaluation metrics (Dice Score, Jaccard Index, Sensitivity, Specificity)

Related Work

Segmentation-Based Methods

Based on the word cloud of paper titles in the medical image analysis community (Jiao et al., 2024), segmentation is one of the most active areas of study and has the highest frequency. The encoder-decoder architecture has gone through numerous variations since the launch of U-Net in Ronneberger et al. (2015) for medical image segmentation to improve it by redesigning skip connections (Ibtehaz and Rahman, 2020) incorporating residual/dense convolution blocks (Peng et al., 2023), attention mechanisms (Huang et al., 2020), and much more. Rathee et al. (2024) analyzed how different fuzzy clustering methods perform on segmenting noisy brain tumor images from CT and MRI scans. It points out that traditional distance metrics often have their drawbacks, but the Manhattan distance metric performs better. It not only yields better segmentation results but also operates more efficiently, even in the presence of image noise. Liu et al. (2024) introduced a new method for segmenting brain tumors that incorporates a Cross-Modal Attention mechanism into a lightweight de-noising network. This approach utilizes multi-sequence MRI images to enhance segmentation accuracy while maintaining low computational requirements. Tested on the BraTS2023 dataset, the model delivers improved performance, particularly in terms of Hausdorff distance, and achieves competitive Dice scores with faster inference times on edge devices. It also addresses the limitations of manual segmentation and introduces EDB-Diff, a streamlined diffusion model that enhances feature extraction and integration. The inference time is sped up, and the number of parameters is reduced drastically by this model. Mansur et al. (2024) explored methods for segmenting brain tumors using MRI scans, with a focus on gliomas, utilizing a Kaggle dataset. It discusses various segmentation techniques, including threshold-based, region-based, and U-Net-based methods, and highlights issues such as data quality and class imbalance. The Dice Similarity Coefficient, which measures the effectiveness of different approaches, indicates that threshold-based techniques achieved the highest accuracy, while the U-Net struggled with overfitting and model complexity. The work emphasizes the need for accurate tumor segmentation for medical diagnosis and recommends future research to enhance U-Net performance and explore hybrid techniques.

Alsenan et al. (2022) discussed a novel deep learning architecture, known as MobileUNetV3, which utilizes MobileNetV3 large as the encoder and Unet as the decoder for segmenting spinal cord gray matter (SCGM). The model addresses challenges in image size variability by standardizing image size to 224×224 pixels and utilizes MobileNetV3 Large efficient feature extraction capabilities. A dataset of 80 healthy patients with 1,092 MRI images was used to train and test MobileUNetV3, producing lofty performance metrics, including a Dice similarity coefficient of 0.87 and a Jaccard index score of 0.78. The study also demonstrates how the efficiency of the model is affected by batch sizes and optimizers, demonstrating that the best results were obtained using a batch size of 8 and the RMSProp optimizer. MobileNetV2 and U-Net are combined to create the RMU-Net model by Saeed et al. (2021), which is used for segmenting brain tumors. U-Net designs utilize U-Net as the decoder and MobileNetV2 as the encoder to achieve high Dice scores on the BraTS 2020 datasets. The research indicates the importance of efficient segmentation methods for improving patient outcomes and medical diagnosis. Additionally, metrics such as the Dice Similarity Coefficient are used to evaluate the accuracy of segmentation. In Popat et al. (2023), the Box-Unet architecture, a brain tumor segmentation model, is introduced. It is built upon the classic U-Net architecture. It is designed to improve accuracy and efficiency in MRI-based tumor detection. The key innovation lies in adding "boxes", sets of upsampling and convolution layers, in the expansive path, which capture tumor features in a better way. The model was trained on the BraTS 2020 dataset. By using an Adam optimizer, dynamic learning rate, and categorical cross-entropy loss, the model outperforms the conventional U-Net in terms of sensitivity, accuracy, and a 1.56% increase in the Jaccard coefficient. The results show that Box-Unet performs better than conventional models, making it a promising tool to help radiologists. To further improve performance, Attention modules can be added in the future inside Box-Unet.

Segmentation and Classification-Based Methods

The importance of a segmentation-based model, which first defines tumor sub-regions before using them to classify or predict survival, for brain tumor prognosis is still being emphasized by recent studies. In order to classify molecular subtypes, Sun et al. (2024) segmented glioma subregions and extracted radiomics characteristics, achieving robust performance in subtype classification. Similar to this, Mahmoudi et al. (2024) presented a radio genomic framework in which radiomics descriptors from segmented tumor regions were combined with genomic and clinical indicators to improve the prediction of overall survival in patients with glioblastoma.

Liu et al. (2024) expanded the segmentation-classification method beyond adult gliomas to children's diffuse midline gliomas, using machine learning and sub-region characteristics to allow for early survival predictions. Another development in deep learning is segmentation-guided learning. Kwon et al. (2024) designed a network that specifically uses tumor masks to focus on feature learning for OS classification. More recently Wan et al. (2025) compared radiomics-based classifiers with deep-learning models on recurrent high-grade gliomas, underscoring the complementary strengths of handcrafted and learned features. Collectively, these works illustrate that segmentation remains a vital first step for interpretable and accurate classification-driven predictions in brain tumor analysis.

Explainability-Based Methods

In 2024-25, the use of Explainable Artificial Intelligence (XAI) in brain tumor segmentation has accelerated, with more focus on improving clinical trust and interpretability. The recent research shows the increasing usage of attribution-based techniques such as Grad-CAM, saliency maps, and influence functions to enhance transparency in deep learning outputs. For instance, Farhan et al. (2025) proposed an ensemble dual-modality framework (XAI-MRI) where Grad-CAM heatmaps were overlaid on tumor masks to make segmentation decisions more interpretable for clinicians. Similarly, Lakshmi et al. (2025) combined U-Net with Bayesian uncertainty estimation, generating not only accurate segmentations but also confidence and explanation maps, addressing the reliability concerns often raised in medical deployment.

More Advanced strategies extend beyond saliency visualizations. Torda et al. (2025) investigated influence-based Explainability, offering deeper insights into which training examples most affect segmentation outcomes. Transformer-based approaches, such as the TransXAI framework by Zeineldin et al. (2024), leverage hybrid CNN–CNN-Transformer architectures with Grad-CAM explanations, demonstrating that high segmentation accuracy can be paired with interpretable attention-driven visualizations. Meanwhile, web-deployed frameworks (Aksoy et al., 2025) and Neuro-XAI (Saeed et al., 2024) illustrate how explainable segmentation can be integrated into interactive clinical systems. Collectively, these works indicate a clear trend toward embedding XAI modules directly within segmentation pipelines, bridging the gap between algorithmic performance and clinical adoption.

Research Gap

Though the architecture in the literature survey mentioned above demonstrates progress in efficient, accurate segmentation of brain tumors, some limitation exists in the current research, which need to be addressed as follows:

- Many existing models give high precision for brain tumor segmentation, but require significant computing power and time. Therefore, it is not easy to use them in hospital settings
- It is still difficult to precisely identify all tumor parts, such as the enhancing tumor and the tumor core. Even with good models, accurately identifying the exact borders of each tumor sub-region remains a challenge
- While various U-Net-like architectures exist, a persistent challenge remains in developing a model that generalizes to the variability and noise found in diverse multi-modal MRI scans

Materials and Methods

BraTS 2020 Datasets

The model proposed in this paper was assessed and trained on the well-established benchmark for brain tumor segmentation, the BraTS 2020 dataset. This dataset is publicly available on Kaggle. The dataset contains four multi-modal MRI scans of brain tumors and their corresponding ground truth. It includes a total of 369 subjects in the training set. The size of all the MRI images in the dataset is 240×240×155 (Usman Akbar et al., 2024). The four multi-modal MRI scans are explained as follows:

- FLAIR uses some inversion recovery methods and specialized pulse sequences to decrease the cerebrospinal fluid signal. This method detects the water content in the brain's tissues (Lin and Lin, 2024)
- T1 imaging utilizes the weak signal intensity from tissues filled with liquid and the strong signal intensity from tissues containing fat to create contrast-rich images of brain regions (Lin et al., 2024)
- By exploiting the strong signal from tissues containing fat and the low signal from fluid tissues, T1ce scans provide high-contrast pictures of brain regions (Lin et al., 2024)
- T2 imaging measures the amount of water in brain regions by using a strong signal from fluid tissues (Lin et al., 2024)
- The accurately annotated brain tumor regions, as determined by experts, are represented by the ground truth (Lin et al., 2024)

The different MRI modalities provide an essential understanding of brain tumor structures and abnormalities. This comprehensive understanding of the MRI data significantly contributes to the work of brain tumor segmentation (Lin et al., 2024). Figure 1 illustrates sample MRI images of the BraTS 2020 Datasets of all four modalities.

Data Preprocessing

To improve the model's performance and consistency, preprocessing procedures were applied to the raw MRI scans of the BraTS 2020 dataset, as shown in Figure 2. The preprocessing steps included:

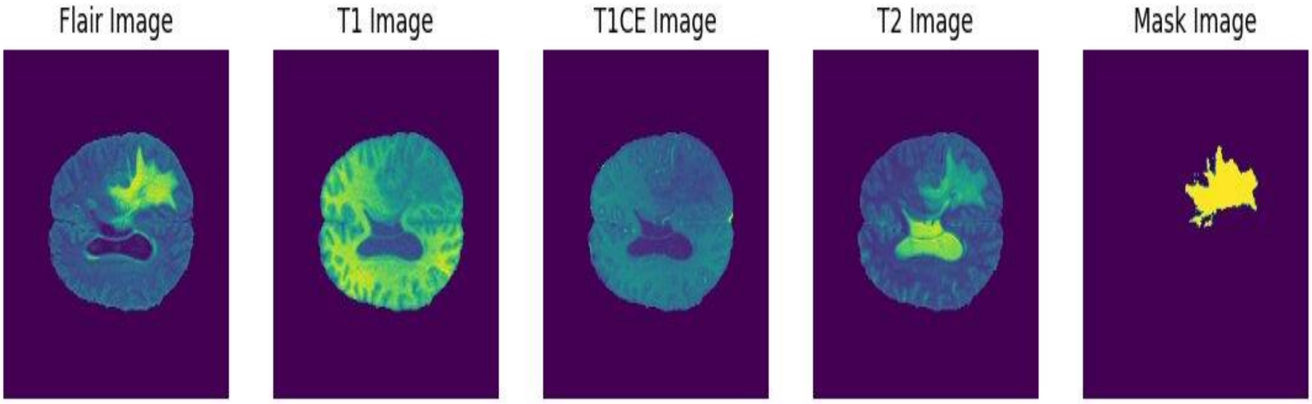

Fig. 1: Visualization of Sample MRI Images of BraTS 2020 datasets

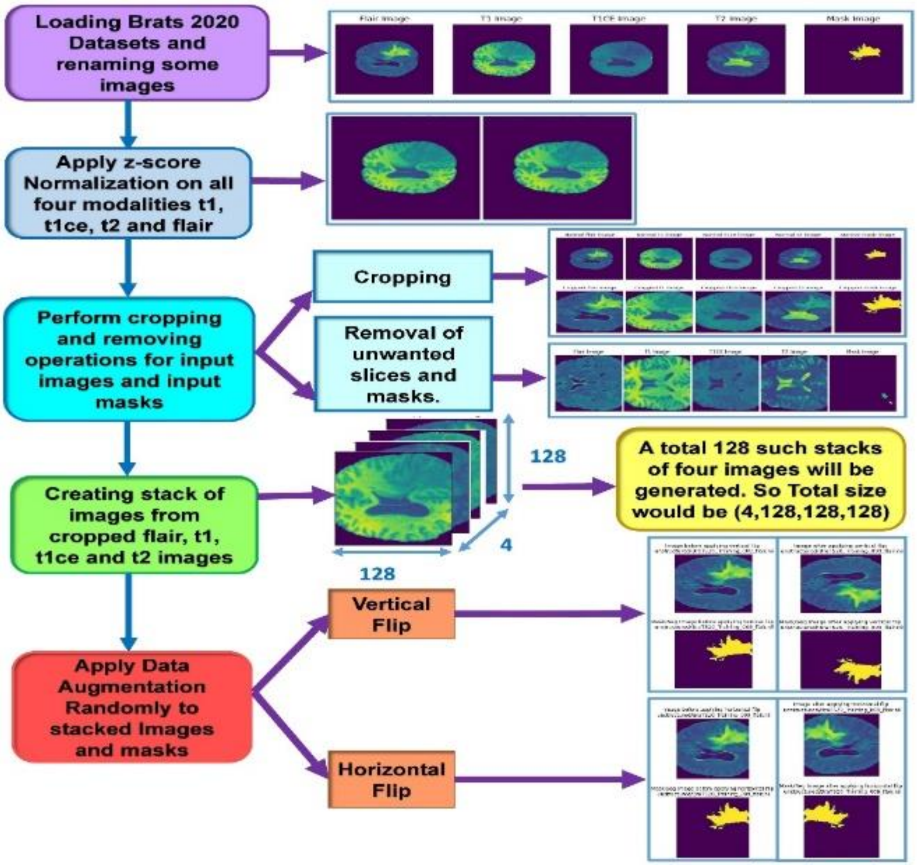
Loading Brats 2020 Datasets and renaming some images
Flair image
T1 image
T1CE image
T2 image
Mask image
Apply z-score Normalization on all four modalities t1, t1ce, t2 and flair
Perform cropping and removing operations for input images and input masks
Cropping
Removal of unwanted slices and masks.
Creating stack of images from cropped flair, t1, t1ce and t2 images
128
128
4
A total 128 such stacks of four images will be generated. So Total size would be (4,128,128,128)
Apply Data Augmentation Randomly to stacked Images and masks
Vertical Flip
Horizontal Flip

Fig. 2: Flowchart of Preprocessing steps applied to BraTS 2020 datasets

- Renaming some images for better regex search at the time of image loading of the BraTS 2020 dataset
- Z-score normalization is applied to T2, FLAIR, T1, and T1ce, as shown in Figure 3

All of the Bra TS 2020 datasets contain MRI scans with pixel intensity levels ranging from 0.0 to 583.0. After Z-score normalization, the range of values for pixel intensity of MRI images was -0.569 to 2.962. Equation (1) shows the formula for Z-score Normalization:

$$Z = \frac{X-\mu}{\sigma} \tag{1}$$

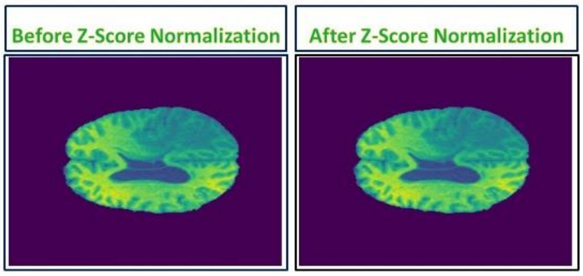

Fig. 3: Z-Score Normalization of Sample MRI Image of BraTS 2020 dataset

123Where X ground truth reflects the original pixel value, μ denotes the mean of the image pixel values, σ ground truth reflects the standard deviation of the image pixel values, and Z denotes the normalized values:

- Performing the following operations for input images and input masks:

a. Cropping: All images are cropped and resized from 240×240×155 to a size of 128×128×128
b. Removal of unwanted slices: Those slices with a tumor area of less than 1% are removed

- Creating stack images from T2, T1ce, FLAIR, and T1 images, excluding the mask
- We used standard methods of data augmentation, like random vertical and horizontal flipping of stacked images

These preprocessing steps offer several advantages. They standardize and refine raw MRI data, which helps improve models' performance and robustness. Z-score Normalization and Image cropping are important for consistent input dimensions and intensity scale, which are important for stable deep learning training and effective feature learning. Stacking multi-modal images and removing irrelevant slices optimizes the dataset for tumor-specific learning. There were around 25 images in the BraTS 2020 dataset where the tumor area was less than 1% of the entire slice area. Some of these slices contained no visible tumor or corresponding ground truth, while others lacked both brain images and masks altogether. Such slices were removed during pre-processing to ensure that only informative samples contributed to training and evaluation. Data augmentation through random flipping expands the training data, leading to a more accurate model that generalizes well to unseen patient scans

Experimental Setup and Hyperparameter Tuning

The tests were conducted using a 16GB GPU RAM and an NVIDIA RTX A4000 GPU. The models were developed utilizing the deep learning framework PyTorch 2.3.1 along with the PyTorch Lightning framework 2.2.4. Python version 3.12.0 was used. The CUDA 12.2 software interface was used for parallel processing. Hyperparameters are configurable settings that control the training process of the deep learning model. They influence the model's ability to learn effectively and achieve optimal segmentation performance for brain tumors. Table 1 presents the various Hyperparameter values used in the experiment for model training

The suggested Two-Stage Cascaded MNetV3UNet architecture demonstrated encouraging outcomes on the Brats 2020 dataset. The dataset was split into training, validation, and test sets with a 70/20/10 ratio at the slice level. With the available dataset size, performing a subject-level split would have significantly reduced the number of training samples, leading to underfitting, especially for deep architectures. The slice-level split provided a much larger and more diverse sample pool for training. Several brain tumor segmentation studies using 2D slice-based approaches have adopted slice-level splitting due to the same trade-off between subject count and sample size. While subject-level splitting is ideal, slice-level splitting remains a widely reported practice in 2D deep learning pipelines

The number of samples in the train, validation, and test sets and the total count of images are displayed in Table 2.

Table 1: Hyperparameters used for Proposed Architecture

Model Training Hyperparameter	Values
Learning Rate	0.001
Optimizer	[Adam, AdamW, RMSProp, SGD]
Loss	Categorical Cross-Entropy
Epochs	135
Batch Size	16
Input Size	128 × 128 × 4
Output Size	128 × 128 × 4

Table 2: Number of Images in the Train Set, Validation Set, and Test Set

Type of datasets	Percentage	Number of Samples	Total Images
Training Set	70%	33047	47207
Validation Set	20%	9440	
Test Set	10%	4720	

Proposed Architecture (Two-Stage Cascaded MNetV3Unet)

A Two-Stage cascaded deep learning model, MNetV3UNet, is designed for effective brain tumor segmentation. It first integrates the effective MobileNetV3 large, which is utilized in the contracting path, and the U-Net architecture, which is utilized in the expanding path. In the second stage, it again utilizes MobileNetV3 large at the encoder position, paired with the U-Net architecture at the decoder position. In both stages, the U-Net encoder is replaced with MobileNetV3 large, whereas the decoder for U-Net remains unchanged.

In the first stage, the network generates coarse multi-class probability maps for the whole tumor, tumor core, and enhancing tumor. These soft probability maps are directly passed as an input to the second-stage network, rather than being thresholded into binary masks or concatenated with the original MRI modalities. This inter-stage connection allows the refinement stage to operate on the continuous probability distributions, preserving uncertainty information from Stage 1. The refinement mechanism works by learning to identify and correct spatial inconsistencies, blurred boundaries, and false positives in the coarse maps. In particular, the second-stage network sharpens boundary delineation by focusing on regions where Stage 1 exhibits low confidence (values between 0 and 1), while reinforcing confident predictions. Thus, Stage 1 serves as a coarse locator of tumor regions, and Stage 2 acts as a corrective module that produces sharper, more reliable tumor boundaries and improved sub-region separation. The diagram in Figure 4 shows the overall pipeline of the proposed architecture. The following sections contain a description of the proposed architecture.

Input and Initial Convolution (Stage 1 Encoder)

The proposed architecture processes multi-modal MRI slice $I^{(1)} \in \mathbb{R}^{H \times W \times C_{in}}$, where $I^{(1)}$ denotes the input layer at stage 1, $H = 128$ and $W = 128$ are the height and width of the MRI slice, and $C_{in} = 4$ is the number of input channels, which represents FLAIR, T1, T1ce, and T2 modalities. The first layer applies a 2D Convolution $Conv_{initial}$ with a 3 × 3 kernel and a stride of 2, followed by Batch Normalization (BN) and an activation function called hard-swish (h-swish).

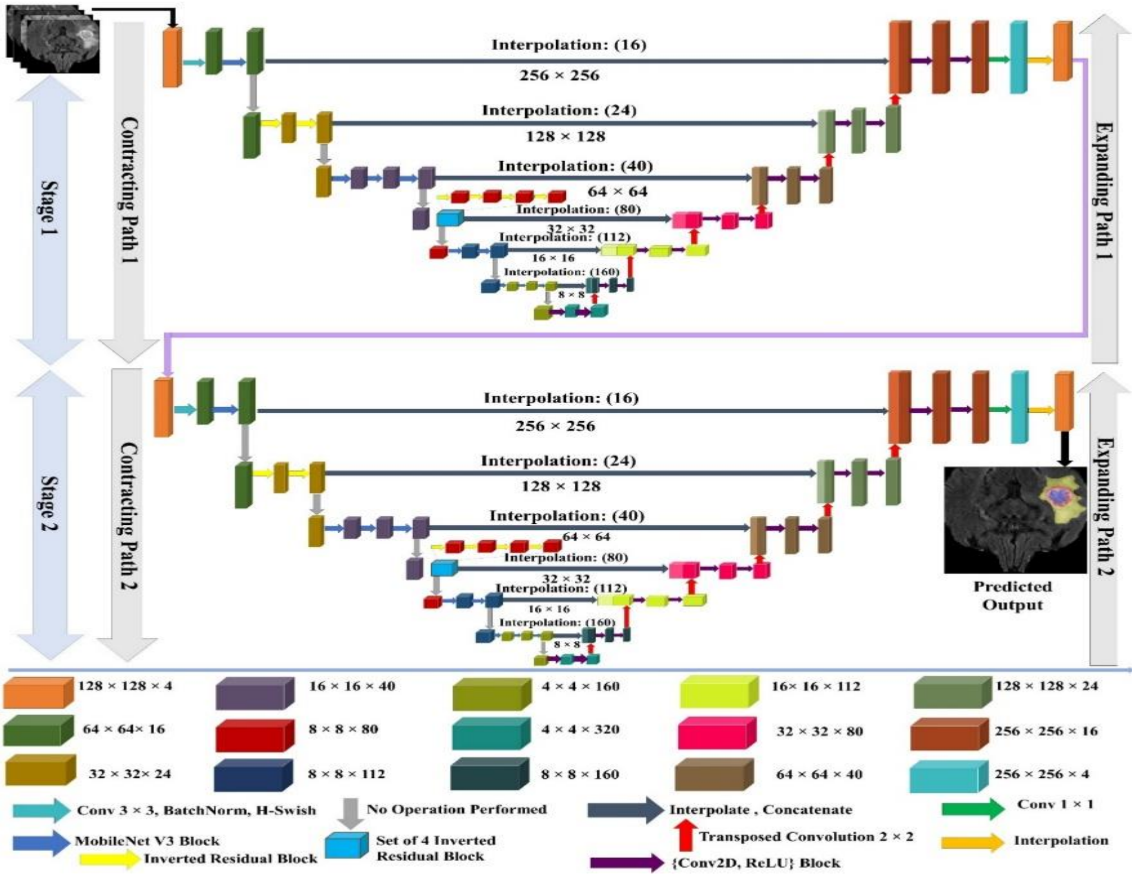

Fig. 4: The Proposed Architecture: Two-Stage Cascaded MNetV3UNet

Let the kernel be $k_{init} \in \mathbb{R}^{3\times3\times4\times16}$. The output feature map $F_0^{(1)} \in \mathbb{R}^{64\times64\times16}$ is computed as shown in Equation 2:

$$F_0^{(1)} = h - swish\left(BN\left(I^{(1)} * k_{init}\right)\right) \in \mathbb{R}^{64\times64\times16} \qquad (2)$$

Where * denotes 2D convolution operation. It halves the spatial dimensions from (128×128) to (64×64) and increases the number of channels to 16.
The h-swish activation function is given by the following Equation 3:

$$h - swish(x) = \frac{x \cdot ReLU6(x+3)}{6} \qquad (3)$$

Where ReLU6(x) is an alteration of the Rectified Linear Unit. Here, the maximum size of the activation is 6 (Alsenan et al., 2022).

MobileNetV3 Blocks and Inverted Residual Blocks (IRB) Used in the Encoder (Stage 1)

The encoder's primary role is to extract increasingly complex features from the input images while gradually reducing their spatial dimensions (height and width) and increasing the number of feature channels (depth). The core of the encoder consists of several "Down-Sampling (DS) layers" (DS1 through DS6). Each of these layers (except DS5, which only increases the number of channels) further reduces spatial dimensions (e.g., 64×64→32×32→16×16→8×8→4×4) while increasing the number of feature channels (e.g., 16→24→40→80→112→160).

These down-sampling layers heavily utilize IRB and MobileNetV3 blocks (IRB + Squeeze and Excitation (SE)). In DS1, one set of MobileNetV3 blocks is used with the ReLU activation function. In DS2, two sets of IRB Blocks with ReLU activation functions are used. In DS3, three sets of MobileNetV3 blocks are used with the ReLU activation function. In DS4, four sets of IRB Blocks with h-swish activation function are used. In DS5, two sets of MobileNetV3 blocks are used with h-swish activation function, and in DS6, three sets of MobileNetV3 blocks are used with h-swish function. MobileNetV3 blocks are renowned for their computational efficiency and robust feature extraction capabilities, making them well-suited for complex tasks without incurring excessive computational overhead. The MobileNetV3 block is shown in Figure 5. A single MobileNetV3 block with input $X \in \mathbb{R}^{H\times W\times C_{in}}$ can be described as follows.

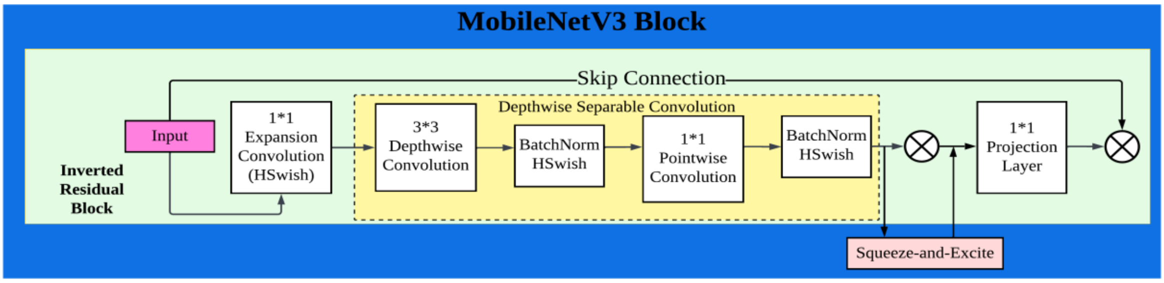

Fig. 5: MobileNetV3 block used in the Proposed Architecture (Dahou et al., 2023)

Pointwise Expansion Convolution

The first step is a 1 × 1 convolution that expands the channel dimension. Let f_{exp} be the output of this operation, as shown by Equation (4):

$$f_{exp} = Act\left(BN\left(X * k_{exp}\right)\right) \quad (4)$$

Where $k_{exp} \in \mathbb{R}^{1 \times 1 \times c_{in} \times c_{exp}}$ and c_{exp} is the expanded channel dimension. Act means activation function, which can be either the ReLU activation function or the h-swish activation function, depending on the block's configuration. For example, in DS1, DS2, and DS3 layers, the ReLU activation function is applied, and in DS4, DS5, and DS6 layers, the h-swish activation function is used.

Depthwise Convolution

After the expansion, each expanded channel undergoes separate spatial filtering using Depthwise convolution. Each of the c_{exp} channels is subjected to a different $k \times k$ filter in this procedure. Depthwise Separable Convolution, denoted by f_{dsc}, can be given by the following Equation:

$$f_{dc} = Act\left(BN\left(f_{exp} \circledast k_{dc}\right)\right) \quad (5)$$

Where $k_{dc} \in \mathbb{R}^{k \times k \times c_{exp} \times 1}$ and $\circledast$ denotes Depthwise Convolution.

Squeeze-and-Excitation (SE) Block (Hu et al., 2018)

Applied to the output of the Depthwise convolution, the SE block dynamically recalibrates channel-wise feature responses. The squeeze and excitation block is shown in Figure 6.

Squeeze (Global Average Pooling): Spatial information is aggregated into a channel descriptor. For $f_{dc} \in \mathbb{R}^{H' \times W' \times c_{exp}}$:

$$Z = GAP(f_{dc}) = \frac{1}{H'\,W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} (f_{dc})_{i,j} \quad (6)$$

Where, $Z \in \mathbb{R}^{1 \times 1 \times c_{exp}}$.

Excitation (Two fully Connected Layers): Two 1 × 1 convolutions acting as FC layers with non-linearities predict channel-wise scaling factors:

$$S = \sigma\left(FC_2\left(ReLU\left(FC_1(Z)\right)\right)\right) \quad (7)$$

Where FC_1 reduces channels by a ratio r = 4 and FC_2 restores them to c_{exp}. σ is the sigmoid activation. $S \in \mathbb{R}^{1 \times 1 \times c_{exp}}$

Scale: The Depthwise convolution result is multiplied element-wise by the scaling factors to produce the SE block's output, f_{se} :

$$f_{se} = f_{dc} \odot S \quad (8)$$

Where $\odot$ represents element-wise multiplication.

Pointwise Projection Convolution: The channel dimension is finally reduced to C_{out}:

$$f_{proj} = BN(f_{se} * \ k_{proj}) \quad (9)$$

Where $k_{proj} \in \mathbb{R}^{1 \times 1 \times c_{exp} \times c_{out}}$.

Residual Connection

A residual connection is added if the input X and the output f_{proj} have the same spatial dimensions and the same channel dimensions (c_{in}, c_{out}). Otherwise, the residual connection is omitted, and the output is $= f_{proj}$:

$$Y = \begin{cases} f_{proj} + X, & \text{if dimensions match} \\ f_{proj} \quad , & otherwise \end{cases} \quad (10)$$

The MobileNetV3 block, with its Inverted bottleneck, Depthwise separable convolution, and squeeze-and-excitation module, enables the encoder to extract rich, multi-scale features while maintaining computational efficiency.

The Bottleneck layer has two sequential convolutional blocks. Each of these blocks is composed of two 3 × 3 2D convolution layers, each followed by a ReLU activation function. The first convolutional block transforms the feature map from 160 channels to 320 channels. The second block refines these 320-channel features. The final output of the bottleneck layer is $B^{(1)} \in \mathbb{R}^{4 \times 4 \times 320}$.

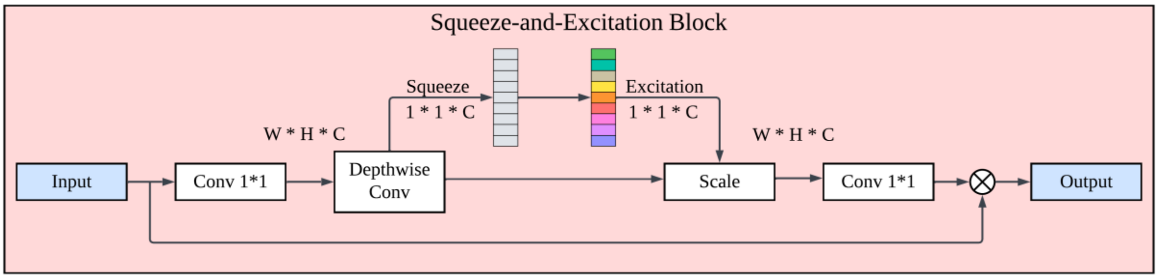

Fig. 6: Squeeze and Excitation block used inside MobileNetV3 Block (Huet al., 2018)

Decoder (Stage 1)

Each stage of the decoder begins with upsampling via transposed convolution, which increases the spatial dimensions of the feature map using a 2 × 2 kernel and a stride of 2. The crucial part of UNet architecture is the skip connection. The encoder feature map (E_j) is first interpolated using bilinear interpolation to match the channel count of the Upsampled Feature map U. The upsampled feature map U is then concatenated with the corresponding more spatially detailed feature map (E_j) from the encoder. After concatenation, refinement blocks are applied. These consist of two sequential 3 × 3 2D convolutional layers, each followed by a ReLU activation function. This iterative process of upsampling, interpolating, concatenating, and refining continues through multiple decoder stages. It gradually increases the spatial resolution until the final segmentation map reaches a resolution of 256 × 256 pixels ($D_{1,6}$).

Final Convolution and Interpolation (Stage 1 Output)

After the final decoder block, a 1 × 1 convolution is applied to the refined feature map $D_{1,6}$ to generate a preliminary segmentation map $S^{(1)} \in \mathbb{R}^{256 \times 256 \times N_{classes}}$, where $N_{classes}$ represents the number of target classes. Here, $N_{classes}$ represents the number of target classes and is equal to 4 (Three tumor sub-regions, ET, TC, and WT, and background):

$$S^{(1)} = conv_{final}(D_{1,6}) \in \mathbb{R}^{256 \times 256 \times N_{Classes}} \tag{11}$$

Bottleneck Layer (Stage 1 Encoder)

Finally, a bilinear interpolation operation is applied to resize the segmentation map to the original input size:

$$O^{(1)} = \text{Interpolate}(S^{(1)}, 128 \times 128 \times N_{Classes}) \tag{12}$$

This is the output of the first stage.

Stage 2: The second stage is exactly the same as the first stage. It takes the output of the first stage, $O^{(1)} \in \mathbb{R}^{128 \times 128 \times N_{Classes}}$, as its input, $I^{(2)} = O^{(1)} \in \mathbb{R}^{128 \times 128 \times N_{Classes}}$. It processes this input through the same sequence of MobileNetV3 and IRB blocks in the encoder, transposed convolutions and interpolated skip connections in the decoder, and a final 1 × 1 convolution and bilinear interpolation to produce the final refined segmentation map $O^{(2)} \in \mathbb{R}^{128 \times 128 \times N_{Classes}}$.

Feature Extraction and Refinement

In the proposed cascaded MobileNetV3–UNet, feature extraction is performed using an encoder–decoder pipeline with MobileNetV3 inverted residual blocks and Squeeze-and-Excitation (SE) modules to capture both local and global features. The first stage produces a coarse tumor map, which is concatenated with the MRI channels and re-encoded in the second stage. This inter-stage transfer refines boundaries and improves detection of small or ambiguous lesions, resulting in consistently higher Dice and Jaccard scores than baseline models (Tables 7–9).

The refinement process specifically addresses three major challenges in MRI tumor segmentation:

- Small and irregular tumours: The cascaded refinement ensures missed or fragmented tumor regions from the coarse segmentation are recovered by leveraging the multi-scale receptive fields in MobileNetV3 inverted residual blocks
- Intensity similarity with surrounding tissues: Squeeze-and-Excitation (SE) blocks recalibrate channel-wise feature maps, enhancing tumor-relevant features while suppressing background regions of similar intensity
- Boundary delineation: Skip connections and concatenation of coarse predictions guide the second stage to refine ambiguous boundaries and yield sharper delineation

Results

The study examined the execution of the architecture using various kinds of quantitative metrics, including the

Dice score, Jaccard Index, Sensitivity, and Specificity. Dice score measures the overlap between the predicted result and the ground truth and is a broadly used metric for Brain Tumor Segmentation. The Dice score calculated for each tumor structure: WT, ET, and TC, is given by the following Equation (13):

$$DiceScore(P,G) = 2 \times \frac{|P \cap G|}{|P|+|G|} \tag{13}$$

Where P is the predicted segmentation result, and G is the Ground Truth. The Jaccard Index, also known as IoU, measures the similarity between the predicted and ground truth segmentation by dividing the intersection by their Union and is calculated by the following Equation (14):

$$Jacaard\ Index(P,G) = \frac{|P \cap G|}{|P \cup G|} \tag{14}$$

Sensitivity measures the proportion of actual positive tumor pixels that the model has correctly identified:

$$Sensitivity = \frac{TrPos}{TrPos+FaNeg} \tag{15}$$

Where $TrPos$, $FaNeg$ represent True Positive and False Negative, respectively. Specificity measures the proportion of actual negative (non-tumor) pixels that were correctly identified by the model:

$$Specificity\ \frac{TrNeg}{TrNeg+FaPos} \tag{16}$$

Where $TrNeg$ represents True Negative and $FaPos$ represents False Positive.

The proposed architecture was trained on various optimizers: Adam, AdamW, RMS Prop, and SGD. The role of the optimizer is to minimize the loss function by adjusting the model's internal weights and biases during training. Out of all the optimizers, Adam gave the best performance while RMS Prop gave the worst performance. The results presented in Tables 3 and 4 demonstrate the performance of the proposed Cascaded MNetV3UNet architecture on validation datasets using various optimizers. This is important because the validation dataset shows more realistic information about how the model generalizes to unseen data. The proposed model with the RMSProp optimizer shows the weakest performance across all metrics and tumor sub-regions compared to other optimizers. While its specificity for WT and TC is 100% which indicates an extremely cautious and overly conservative model, so cautious that it segments very little, thus avoiding false positives, but possibly missing many true positives (low sensitivity) or under-segmenting the actual tumor (low Dice/Jaccard). This optimizer clearly struggles to generalize. SGD demonstrated a substantial improvement over RMSProp, providing robust and consistent segmentation performance across all metrics and tumor sub-regions, indicating decent generalization. AdamW achieved significantly higher Dice and Jaccard scores than SGD, showcasing excellent segmentation quality and strong generalization across tumor sub-regions on unseen data. The proposed model achieves the greatest overall performance using the Adam optimizer. It obtains the highest Dice score of 88.35 for WT, 89.03 for TC, and 92.30% for ET. The efficient combination of MobileNetV3-Large's feature extraction capabilities and U-Net's encoder-decoder structure is responsible for the suggested architecture's high performance. The proposed model with SGD performs better than RMSProp.

Table 5 shows the number of hours required by the proposed architecture for each optimizer. It is clear that the proposed architecture with RMSProp took the minimum time, and the proposed architecture with AdamW optimizer took the maximum time for training.

Ablation Study

To evaluate the contribution of different architectural components, an ablation study was conducted across nine different architectures.

Table 3: Comparison of the Proposed Architecture with Different Optimizers on BraTS 2020 Validation Set (Dice Score & Jaccard Score**)**

Optimizer	Dice Score (%)			Jaccard Score (%)		
	WT	TC	ET	WT	TC	ET
RMSProp	48.34	67.68	71.15	48.34	67.68	70.60
SGD	72.99	75.10	82.36	67.24	71.83	79.08
AdamW	87.59	88.57	92.10	82.90	85.67	89.50
Adam	88.35	89.03	92.30	83.76	86.15	89.68

Table 4: Comparison of the Proposed Architecture with Different Optimizers on BraTS 2020 Validation Set (Specificity & Sensitivity**)**

Optimizer	Specificity (%)			Sensitivity (%)		
	WT	TC	ET	WT	TC	ET
RMSProp	100	100	87.10	48.34	67.68	71.79
SGD	99.29	99.74	95.82	79.18	81.03	87.32
AdamW	99.70	99.88	97.84	88.35	89.17	92.82
Adam	99.72	99.89	97.96	88.91	89.37	92.88

Table 5: Comparison of Training Time for the Proposed Model with Different Optimizers (135 Epochs, Learning Rate 0.001, Batch Size 16**)**

Architecture Information	Training Time
Proposed architecture with Adam Optimizer	14h 44m 20s
Proposed architecture with AdamW Optimizer	16h 4m 20s
Proposed architecture with RMSProp Optimizer	13h 49m 32s
Proposed Architecture with SGD Optimizer	15h 4m 7s

These include the baseline U-Net, Box U-Net, Mobile Net-based U-Nets (V1, V2, V3 Small, V3 Large), MobileNetV3 models without the Squeeze-and-Excitation (SE) block, and the final cascaded MobileNetV3-Large U-Net. Table 6 summarizes the average Dice and Jaccard scores for tumor sub-regions (WT, TC, ET) on the BraTS 2020 validation set. Results clearly show that while baseline U-Net and Box U-Net provide reasonable segmentation, incorporating lightweight Mobile Net encoders significantly improves dice score and Jaccard score. Furthermore, removing the SE block reduces performance, confirming its importance. Finally, cascading MobileNetV3-Large U-Net yields the highest Dice and Jaccard scores, establishing the effectiveness of the proposed design.

The ablation results in Table 6 provide insights into the effectiveness of different design choices in brain tumor segmentation. The baseline U-Net achieves moderate performance across tumor subregions, demonstrating the fundamental strength of the encoder–decoder architecture. Box U-Net slightly improves the results by better handling spatial information, but its gain remains limited. When MobileNet encoders are incorporated, a substantial performance improvement is observed. MobileNetV1 + U-Net and MobileNetV2 + U-Net show clear gains in Dice and Jaccard scores compared to baseline models, highlighting the advantage of using lightweight but expressive encoders. Among them, MobileNetV2 demonstrates a stronger feature extraction capability than MobileNetV1. MobileNetV3-based architectures further enhance segmentation performance, with the Large variant consistently outperforming the Small one due to its greater representational power. The ablation study also reveals that removing the Squeeze-and-Excitation (SE) block significantly decreases accuracy, confirming the importance of channel attention in capturing discriminative tumor features. Finally, the proposed Cascaded MobileNetV3 Large + U-Net achieves the best overall performance across WT, TC, and ET regions. Cascading enables the second stage to refine coarse predictions from the first stage, leading to improved boundary precision and robust tumor delineation. This confirms that both architectural refinements and cascading contribute meaningfully to segmentation quality, with the cascaded model providing the highest Dice and Jaccard scores among all tested variants.

Table 7 presents the model complexity analysis and inference times across all evaluated architectures. The proposed two-stage cascaded MobileNetV3-UNet demonstrated a favorable balance between efficiency and performance, with 11.6M trainable parameters, 11.50 GFLOPs, and an average inference time of 5.75 ms per image. Compared to the baseline U-Net (31.0M parameters, 34.14 GFLOPs, 17.07 ms), the proposed model reduced parameter count by more than 60% and inference time by over 65%. While lightweight variants such as MobileNetV3Small-UNet achieved faster inference (0.26 ms), they exhibited significantly reduced Dice and Jaccard scores (Table 6), underscoring the superior trade-off achieved by the proposed cascaded MobileNetV3-UNet.

Table 8 shows the five-fold cross-validation of all nine architectures. The proposed two-stage cascaded mobile net v3 -Unet produced the best Dice and Jaccard scores with minimum standard deviation, which implies adequate segmentation and the stability of the network between folds. In further ensuring the soundness of the results, we present the 95% Confidence Intervals (CI) in addition to the mean + ± standard deviation. The developed model obtained smaller CIs than those of other architectures, and this implies more robust generalization. Paired t-tests were used to verify that the gains over the baseline U-Net were significant (p<0.05.

Table 6: Ablation Study – Performance Comparison of Proposed and Baseline Architectures on Validation Set

Model	Avg. Dice (%)	Δ vs U-Net	Avg. Jaccard (%)	Δ vs U-Net
U-Net (Baseline)	88.66	–	85.23	–
Box-U-Net (skip modification) (Popat et al., 2023)	88.62	–0.04	85.17	–0.06
MobileNetV1 + U-Net	89.24	+0.58	85.86	+0.63
MobileNetV2 + U-Net	89.26	+0.60	85.89	+0.66
MobileNetV3 Small + U-Net	88.47	–0.19	84.93	–0.30
MobileNetV3 Large + U-Net	89.43	+0.77	85.28	+0.05
MobileNetV3 Small + U-Net (w/o SE)	88.29	–0.37	84.67	–0.56
MobileNetV3 Large + U-Net (w/o SE)	89.23	+0.57	85.19	–0.04
Proposed: Two Stage Cascaded MobileNetV3 Large + U-Net	89.87	+1.21	86.15	+0.92

Table 7: Model Complexity Analysis and Inference Time of Proposed and Baseline Architecture

Model	Trainable parameters	FLOPs (GFLOPs)	Inference time (ms) per image
U-Net (baseline)	31,032,516	34.14	17.07
Box-Unet	22,021,675	24.23	12.12
MobileNetV1-Unet	108,716,516	71.76	35.88
MobileNetV2-Unet	12,334,524	6.11	3.06
MobileNetV3Small-Unet	1,866,978	0.51	0.26
MobileNetV3Large-Unet	5,809,182	1.92	0.96
MobileNetV3Small-Unet (w/o SE)	1,766,778	0.49	0.25
MobileNetV3Large-Unet (w/o SE)	4,879,474	1.61	0.81
Two-Stage Cascaded MobileNetV3 Large + U-Net (Proposed)	11,618,364	11.50	5.75

Table 8: Five-fold cross-validation results (Dice and Jaccard, mean ± standard deviation across folds, with 95% confidence intervals) and paired t-test p-values compared to baseline U-Net

Model	Dice (Mean ± SD, 95% CI)	Jaccard (Mean ± SD, 95% CI)	p-value (Dice vs U-Net)	p-value (Jaccard vs U-Net)
U-Net (Baseline)	0.867 ± 0.012 (0.861–0.873)	0.819 ± 0.014 (0.811–0.827)	–	–
Box U-Net	0.865 ± 0.013 (0.858–0.872)	0.816 ± 0.015 (0.808–0.824)	0.067	0.072
MobileNetV1 + U-Net	0.873 ± 0.011 (0.867–0.879)	0.825 ± 0.013 (0.818–0.832)	0.052	0.049
MobileNetV2 + U-Net	0.875 ± 0.010 (0.870–0.880)	0.828 ± 0.012 (0.822–0.834)	0.041	0.038
MobileNetV3 Small + U-Net	0.868 ± 0.012 (0.862–0.874)	0.820 ± 0.014 (0.812–0.828)	0.081	0.077
MobileNetV3 Large + U-Net	0.879 ± 0.011 (0.873–0.885)	0.832 ± 0.013 (0.825–0.839)	0.032	0.029
MobileNetV3 Small + U-Net (w/o SE)	0.866 ± 0.013 (0.859–0.873)	0.818 ± 0.015 (0.810–0.826)	0.075	0.071
MobileNetV3 Large + U-Net (w/o SE)	0.877 ± 0.011 (0.871–0.883)	0.830 ± 0.012 (0.824–0.836)	0.035	0.031
Proposed: Two-Stage Cascaded MobileNetV3-UNet	0.883 ± 0.009 (0.879–0.887)	0.835 ± 0.010 (0.830–0.840)	0.018	0.015

Table 9: Comparison of Dice Score of Proposed Architecture with Adam Optimizers on BraTS 2020 Dataset and Various State-of-the-Art Methods

Sr. No.	Architecture Information	WT	ET	TC
1.	Dual Decoder 3D-Unet with Self-Supervised Approach (Simsiam network) (Samarasinghe et al., 2025)	84.6	88.4	83.7
2.	Multi-Modal Fusion Framework (Zhang and Pan, 2025)	90.22	78.30	86.62
3.	GARU-NET (Raza and Hashmi, 2024)	90.8	82.4	86.00
4.	3D Attention Unet (Tassew et al., 2024)	88.92	82.84	86.61
5.	ACMINet (Zhuang et al., 2023)	90.61	81.13	84.70
6.	AD-Net (Peng et al., 2023)	90.00	76.00	80.00
7.	Deep Residual U-Net (DResU-Net) (Raza et al., 2023)	86.60	80.04	83.57
8.	Prior Attention Network (Zhao et al., 2022)	90.00	78.00	83.00
9.	3D AGSE-Vnet (Guan et al., 2022)	85.00	67.00	69.00
10.	2D Deep Residual U-Net (DR-Unet 104) (Colman et al., 2021)	86.73	75.14	79.83
11.	Multi-encoder Architecture (ME-Net) (Zhang et al., 2021)	70.24	73.86	88.26
12.	Lesion encoder with DCNN network (Russo, Liu, and Ieva, 2020)	86.87	78.98	80.66
13.	3D encoder-decoder based V-net model (Ballestar and Vilaplana, 2020)	84.63	62.15	75.26
14.	3D Efficient Embedding Network (Messaoudi et al., 2020)	80.68	69.59	75.20
	Proposed Architecture	88.35	92.30	89.03

Though there are Mobile Net-based architectures (e.g., MobileNetV3 Large + U-Net) that also showed competitive performance, their improvements were less regular and had broader confidence intervals. Notably, the two-stage cascaded design offered statistically significant and reproducible gains, relative to the baseline U-Net as well as the single-stage MobileNetV3 Large + U-Net, which justifies the use of a cascaded design over single-stage or deeper multi-stage designs, which are more likely to add complexity without corresponding gains in performance.

Discussion

The proposed model has been compared with 12 different state-of-the-art techniques. MobileNetV3-Large is an efficient and lightweight convolutional neural network that can effectively extract both low-level and high-level image attributes from medical images (Laibacher et al., 2019). However, because the U-Net design can gather and combine multi-scale information through skip connections, it has been extensively employed in the work of image segmentation in the medical domain (Huang et al., 2020). A coarse-to-fine segmentation technique is made possible by the cascaded design of the proposed architecture, which uses the first U-Net's output as the second U-Net's input (Chen et al., 2024). The first U-Net provides a rough segmentation of the brain tumor, which enables the second U-Net to focus on refining the details within the region of interest (Li et al., 2020). Moreover, the use of Adam optimizer in the training process of the proposed architecture leads to faster convergence compared to other optimizers like AdamW, SGD, and RMSProp (Zohrevand and Imani, 2022). The strong performance by the proposed Cascaded MNetV3UNet architecture on the BraTS 2020 dataset demonstrates its potential for real-world brain tumor image segmentation applications. The lightweight nature of the MobileNetV3-Large, combined with the efficient U-Net, makes the proposed architecture suitable for deployment on mobile and embedded systems, allowing for accessible and scalable brain tumor segmentation solutions (Jiang et al., 2024).

The Dice scores of the proposed Cascaded MNetV3UNet architecture with Adam optimizer surpass a number of state-of-the-art methods using the BraTS 2020 dataset, as seen in Table 6. The proposed method achieves Dice scores of 88.35 for WT, 89.03 for TC, and 92.30% for ET. The values marked in bold indicate the highest value. The proposed architecture shows outstanding results in ET and TC as compared to various other state-of-the-art techniques. From a clinical point of view, ET and TC are much more important as they represent the most aggressive, proliferative, and clinically significant regions. The proposed architecture does not perform well in terms of WT as compared to other architectures.

The Segmentation results of the Proposed Architecture on different optimizers are shown in Figure 7. WT is segmented by light yellow colour, ET is segmented by light pink colour, and TC is segmented by blue colour. It clearly shows that the Proposed Architecture with Adam Optimizer gave the best performance, clearly showing WT, TC, and ET regions. With the RMS Prop optimizer, the proposed architecture performed the worst.

Limitations of this Study

Although a 70/20/10 split was used at the slice level to ensure sufficient training data, this approach may introduce a risk of data leakage if slices from the same subject appear across sets. Future work will adopt subject-level or volumetric splits to further improve generalization and clinical reliability.

The majority of research relies heavily on publicly available datasets like BraTS 2020, which may not capture the full heterogeneity of tumor types, imaging protocols, and patient demographics. This limits the generalizability and external validity of the models in diverse clinical settings.

Clinical Significance and Future Work

The proposed cascaded MobileNetV3–UNet has potential for deployment in real-world clinical settings, as its segmentation outputs can be directly interpreted by radiologists, similar to recent interpretability-focused studies (Dasanayaka et al., 2022). Although most AI-based models remain experimental, validation of this approach with medical experts is feasible through retrospective comparison with expert annotations and prospective assessment in clinical workflows. Such collaboration will be crucial to establishing clinical reliability and supporting translation into practice.

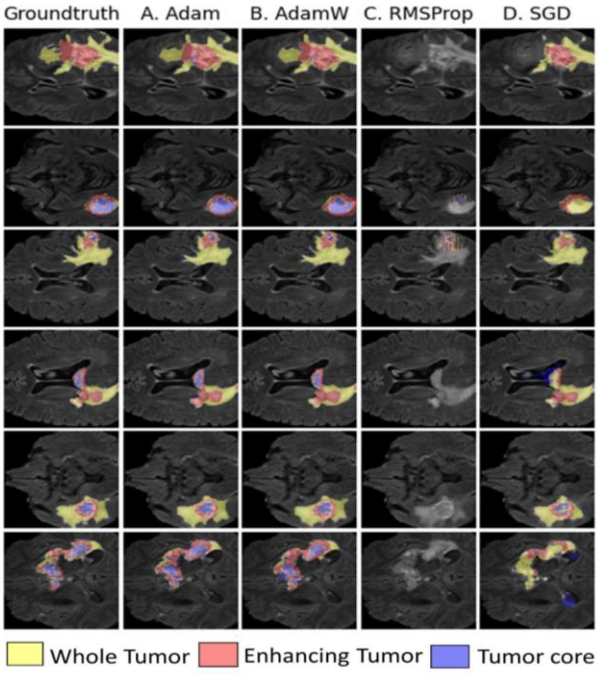

Fig. 7: Segmentation Results on BraTS 2020 Dataset using Proposed Architecture on different optimizers

Conclusion

A coarse-to-fine segmentation strategy is made possible by the architecture's cascaded design. The initial first-stage in this Two-Stage cascaded MNetV3UNet design offers a rough segmentation. The specifics inside the region of interest are then fine-tuned by the second stage. Embedding a MobileNetV3 block into a Two-Stage cascaded architecture is a preferable solution due to its squeeze-and-excitation and inverted residual mechanisms, which it offers, as well as the balance between efficiency and feature extraction. These characteristics enable the model to capture rich, contextually relevant information while maintaining computational efficiency. Additionally, an investigation was conducted to assess the impact of several optimizers, including Adam, AdamW, SGD, and RMSProp, on the performance of the suggested architecture. The findings show that the Adam optimizer produces the quickest convergence among all other optimizers. The suggested Cascaded MNetV3UNet architecture with the Adam optimizer beats several state-of-the-art techniques in ET and TC, according to extensive trials on the BraTS 2020 dataset. Many experiments are required to be performed in the proposed architecture, which can improve the performance of WT. Many other methodologies need to be explored that can be combined with the proposed architecture to enhance its efficiency in terms of WT.

Acknowledgment

The Authors would like to express their deep gratitude to Dr. Trushit Upadhyaya, Dr. Mrugendra Rahevar, and Dr. Spoorthy V for providing all the guidance in writing the research paper. They would also like to thank CHARUSAT - Charotar University of Science and Technology for constant support and encouragement.

Funding Information

The authors have not received any funding.

Authors Contributions

Mayuri Popat: Conceptualization, formal analysis, methodology, writing-original draft, writing-review and edited.

Sanskruti Patel: Methodology, Supervision.

Ethics

The authors declare that the manuscript is an original work and that no ethical issues are associated with this submission. All authors have reviewed and approved the final version of the manuscript.

Reference

Aksoy, S., Demircioglu, P., & Bogrekci, I. (2025). A Web-Deployed, Explainable AI System for Comprehensive Brain Tumor Diagnosis. *Neurology International*, *17*(8), 121. https://doi.org/10.3390/neurolint17080121

Alsenan, A., Ben Youssef, B., & Alhichri, H. (2022). MobileUNetV3—A Combined UNet and MobileNetV3 Architecture for Spinal Cord Gray Matter Segmentation. *Electronics*, *11*(15), 2388. https://doi.org/10.3390/electronics11152388

Amin, B., Samir, R. S., Tarek, Y., Ahmed, M., Ibrahim, R., Ahmed, M., & Hassan, M. (2023). Brain tumor multi-classification and segmentation in MRI images using deep learning. *Image and Video Processing*. https://doi.org/https://doi.org/10.48550/arXiv.2304.10039

Ballestar, L. M., & Vilaplana, V. (2020). Brain Tumor Segmentation using 3D-CNNs with Uncertainty Estimation. *Image and Video Processing*. https://doi.org/https://doi.org/10.48550/arXiv.2009.12188

Chen, B., Sun, Q., Han, Y., Liu, B., Zhang, J., & Zhang, Q. (2024). Adaptive cascaded transformer U-Net for MRI brain tumor segmentation. *Physics in Medicine & Biology*, *69*(11), 115036. https://doi.org/10.1088/1361-6560/ad4081

Colman, J., Zhang, L., Duan, W., & Ye, X. (2021). DR-Unet104 for Multimodal MRI Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 410–419. https://doi.org/10.1007/978-3-030-72087-2_36

Dahou, A., Aseeri, A. O., Mabrouk, A., Ibrahim, R. A., Al-Betar, M. A., & Elaziz, M. A. (2023). Optimal Skin Cancer Detection Model Using Transfer Learning and Dynamic-Opposite Hunger Games Search. *Diagnostics*, *13*(9), 1579. https://doi.org/10.3390/diagnostics13091579

Daimary, D., Bora, M. B., Amitab, K., & Kandar, D. (2020). Brain Tumor Segmentation from MRI Images using Hybrid Convolutional Neural Networks. *Procedia Computer Science*, *167*, 2419–2428. https://doi.org/10.1016/j.procs.2020.03.295

Dasanayaka, S., Shantha, V., Silva, S., Meedeniya, D., & Ambegoda, T. (2022). Interpretable machine learning for brain tumour analysis using MRI and whole slide images. *Software Impacts*, *13*, 100340. https://doi.org/10.1016/j.simpa.2022.100340

Farhan, A. S., Khalid, M., & Manzoor, U. (2025). XAI-MRI: an ensemble dual-modality approach for 3D brain tumor segmentation using magnetic resonance imaging. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1525240

Guan, X., Yang, G., Ye, J., Yang, W., Xu, X., Jiang, W., & Lai, X. (2022). 3D AGSE-VNet: an automatic brain tumor MRI data segmentation framework. *BMC Medical Imaging*, *22*(1), 6. https://doi.org/10.1186/s12880-021-00728-8

Hu, J., Shen, Li, & Sun, G. (2018). Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., & Iwamoto, Y. (2020). UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *Proceeding If the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1059–1055. https://doi.org/10.1109/ICASSP40776.2020.9053405

Ibtehaz, N., & Rahman, M. S. (2020). MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, *121*, 74–87. https://doi.org/10.1016/j.neunet.2019.08.025

Jiang, J., Wang, M., Tian, H., Cheng, L., & Liu, Y. (2024). LV-UNet: A Lightweight and Vanilla Model for Medical Image Segmentation. *Proceeding of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 4240–4246. https://doi.org/10.1109/bibm62325.2024.10822465

Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., & Jin, C. (2024). Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, *169*, 107840. https://doi.org/10.1016/j.compbiomed.2023.107840

Laibacher, T., Weyde, T., & Jalali, S. (2019). M2U-Net: Effective and Efficient Retinal Vessel Segmentation for Real-World Applications. *Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 115–124. https://doi.org/10.1109/cvprw.2019.00020

Lakshmi, K., Amaran, S., Subbulakshmi, G., Padmini, S., Joshi, G. P., & Cho, W. (2025). Explainable artificial intelligence with UNet based segmentation and Bayesian machine learning for classification of brain tumors using MRI images. *Scientific Reports*, *15*(1), 690. https://doi.org/10.1038/s41598-024-84692-7

Li, X., Fang, X., Yang, G., Su, S., Zhu, L., & Yu, Z. (2023). TransU^2-Net: An Effective Medical Image Segmentation Framework Based on Transformer and U^2-Net. *IEEE Journal of Translational Engineering in Health and Medicine*, *11*, 441–450. https://doi.org/10.1109/jtehm.2023.3289990

Li, X., Luo, G., & Wang, K. (2020). Multi-step Cascaded Networks for Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 163–173. https://doi.org/10.1007/978-3-030-46640-4_16

Lin, S.-Y., & Lin, C.-L. (2024). Brain tumor segmentation using U-Net in conjunction with EfficientNet. *PeerJ Computer Science*, *10*, e1754. https://doi.org/10.7717/peerj-cs.1754

Liu, Y., Xie, L., & Ye, W. (2024). EDB-Diff: a EdgeDevice based diffusion network for brain tumor image segmentation. *Multimedia Systems*, *30*(6), 1–342. https://doi.org/10.1007/s00530-024-01580-w

Mansur, Z., Talukdar, J., Singh, T. P., & Kumar, C. J. (2024). Deep Learning-Based Brain Tumor Image Analysis for Segmentation. *SN Computer Science*, *6*(1), 42. https://doi.org/10.1007/s42979-024-03558-x

Messaoudi, H., Belaid, A., Allaoui, M. L., Zetout, A., Allili, M. S., Tliba, S., Salem, Douraied Ben, & Conze, P.-H. (2020). Efficient embedding network for 3D brain tumor segmentation. *ArXiv*, 252–262. https://doi.org/10.1007/978-3-030-72084-1_23

Peng, Y., Chen, D. Z., & Sonka, M. (2023). U-Net V2: Rethinking the Skip Connections of U-Net for Medical Image Segmentation. *Proceeding of the IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5. https://doi.org/10.1109/isbi60581.2025.10980742

Popat, M., Patel, S., Poshiya, Y., & Sai, A. K. (2023). Brain Tumor Image Segmentation using Box-Unet Architecture. *Proceeding of the First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*, 1–7. https://doi.org/10.1109/icaeeci58247.2023.10370887

Raza, A., & Hashmi, M. F. (2024). Multiclass Tumor Segmentation From Brain MRIs Using GARU-Net: Gelu Activated Attention Aware Res-3DUNET for Adaptive Feature Pooling. *IEEE Sensors Letters*, *8*(4), 1–4. https://doi.org/10.1109/lsens.2024.3370974

Raza, R., Ijaz Bajwa, U., Mehmood, Y., Waqas Anwar, M., & Hassan Jamal, M. (2023). dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI. *Biomedical Signal Processing and Control*, *79*, 103861. https://doi.org/10.1016/j.bspc.2022.103861

Rathee, J., Kaur, P., & Singh, A. (2024). Fuzzy Clustering Based Noisy Image Segmentation of MRI/CT Scan Brain Tumor Images Using Different Distance Metrics as Similarity Measure. *SN Computer Science*, *5*(6), 1–5. https://doi.org/10.1007/s42979-024-03102-x

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI* (Vol. 9351, pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Russo, C., Liu, S., & Di Ieva, A. (2022). Spherical coordinates transformation pre-processing in Deep Convolution Neural Networks for brain tumor segmentation in MRI. *Medical & Biological Engineering & Computing*, *60*(1), 121–134. https://doi.org/10.1007/s11517-021-02464-1

Samarasinghe, D., Wickramasinghe, D., Wijerathne, T., Meedeniya, D., & Yogarajah, P. (2025). Brain Tumour Segmentation and Edge Detection Using Self-Supervised Learning. *International Journal of Online and Biomedical Engineering (IJOE)*, *21*(05), 127–141. https://doi.org/10.3991/ijoe.v21i05.53405

Saeed, M. U., Ali, G., Bin, W., Almotiri, S. H., AlGhamdi, M. A., Nagra, A. A., Masood, K., & Amin, R. ul. (2021). RMU-Net: A Novel Residual Mobile U-Net Model for Brain Tumor Segmentation from MR Images. *Electronics*, *10*(16), 1962. https://doi.org/10.3390/electronics10161962

Saeed, T., Khan, M. A., Hamza, A., Shabaz, M., Khan, W. Z., Alhayan, F., Jamel, L., & Baili, J. (2024). Neuro-XAI: Explainable deep learning framework based on deeplabV3+ and bayesian optimization for segmentation and classification of brain tumor in MRI scans. *Journal of Neuroscience Methods*, *410*, 110247. https://doi.org/10.1016/j.jneumeth.2024.110247

Tassew, T., Ashamo, B. A., & Nie, X. (2024). Multimodal MRI brain tumor segmentation using 3D attention UNet with dense encoder blocks and residual decoder blocks. *Multimedia Tools and Applications*, *84*(7), 3611–3633. https://doi.org/10.1007/s11042-024-18942-1

Torda, T., Ciardiello, A., Gargiulo, S., Grillo, G., Scardapane, S., Voena, C., & Giagu, S. (2025). Influence based explainability of brain tumors segmentation in magnetic resonance imaging. *Progress in Artificial Intelligence*. https://doi.org/10.1007/s13748-025-00367-y

Ullah, F., Nadeem, M., Abrar, M., Al-Razgan, M., Alfakih, T., Amin, F., & Salam, A. (2023). Brain Tumor Segmentation from MRI Images Using Handcrafted Convolutional Neural Network. *Diagnostics*, *13*(16), 2650. https://doi.org/10.3390/diagnostics13162650

Usman Akbar, M., Larsson, M., Blystad, I., & Eklund, A. (2024). Brain tumor segmentation using synthetic MR images - A comparison of GANs and diffusion models. *Scientific Data*, *11*(1), 256. https://doi.org/10.1038/s41597-024-03073-x

Wan, Q., Lindsay, C., Zhang, C., Kim, J., Chen, X., Li, J., Huang, R. Y., Reardon, D. A., Young, G. S., & Qin, L. (2025). Comparative analysis of deep learning and radiomic signatures for overall survival prediction in recurrent high-grade glioma treated with immunotherapy. *Cancer Imaging*, *25*(1), 5. https://doi.org/10.1186/s40644-024-00818-0

Yousef, R., Khan, S., Gupta, G., Siddiqui, T., Albahlal, B. M., Alajlan, S. A., & Haq, M. A. (2023). U-Net-Based Models towards Optimal MR Brain Image Segmentation. *Diagnostics*, *13*(9), 1624. https://doi.org/10.3390/diagnostics13091624

Yue, G., Zhuo, G., Zhou, T., Liu, W., Wang, T., & Jiang, Q. (2025). Adaptive Cross-Feature Fusion Network With Inconsistency Guidance for Multi-Modal Brain Tumor Segmentation. *IEEE Journal of Biomedical and Health Informatics*, *29*(5), 3148–3158. https://doi.org/10.1109/jbhi.2023.3347556

Zhang, M., & Pan, K. (2025). A Multi-Modal Fusion Framework for Brain Tumor Segmentation Based on 3D Spatial-Language-Vision Integration and Bidirectional Interactive Attention Mechanism. *Computer Vision and Pattern Recognition*.

Zhang, W., Yang, G., Huang, H., Yang, W., Xu, X., Liu, Y., & Lai, X. (2021). ME-Net: Multi-encoder net framework for brain tumor segmentation. *International Journal of Imaging Systems and Technology*, *31*(4), 1834–1848. https://doi.org/10.1002/ima.22571

Zhao, X., Zhang, P., Song, F., Ma, C., Fan, G., Sun, Y., Feng, Y., & Zhang, G. (2022). Prior Attention Network for Multi-Lesion Segmentation in Medical Images. *IEEE Transactions on Medical Imaging*, *41*(12), 3812–3823. https://doi.org/10.1109/tmi.2022.3197180

Zafar, W., Husnain, G., Iqbal, A., Alzahrani, A. S., Irfan, M. A., Ghadi, Y. Y., AL-Zahrani, M. S., & Naidu, R. S. (2024). Enhanced TumorNet: Leveraging YOLOv8s and U-net for superior brain tumor detection and segmentation utilizing MRI scans. *Results in Engineering*, *24*, 102994. https://doi.org/10.1016/j.rineng.2024.102994

Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., Burgert, O., & Mathis-Ullrich, F. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-54186-7

Zohrevand, A., & Imani, Z. (2022). An Empirical Study of the Performance of Different Optimizers in the Deep Neural Networks. *Proceeding of the International Conference on Machine Vision and Image Processing (MVIP)*, 1–5. https://doi.org/10.1109/mvip53647.2022.9738743

Zhuang, Y., Liu, H., Song, E., & Hung, C.-C. (2023). A 3D Cross-Modality Feature Interaction Network With Volumetric Feature Alignment for Brain Tumor and Tissue Segmentation. *IEEE Journal of Biomedical and Health Informatics*, *27*(1), 75–86. https://doi.org/10.1109/jbhi.2022.3214999