Swin-Conv: A Hybrid Swin Transformer and Convolutional Network for Brain Tumor Segmentation

Khadijah
Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia
khadijah@live.undip.ac.id (corresponding author)

Helmie Arif Wibawa
Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia
helmie.arif@live.undip.ac.id

Ragil Saputra
Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia
ragil.saputra@live.undip.ac.id

Rismiyati
Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia
rismiyati@live.undip.ac.id

Sandy Kurniawan
Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia
sandy@live.undip.ac.id

Mohd Hanafi Bin Ahmad Hijazi
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia
hanafi@ums.edu.my

Received: 6 February 2026 | Revised: 9 April 2026 | Accepted: 19 April 2026

 DOI: https://doi.org/10.48084/etasr.18002

ABSTRACT

Accurate segmentation of brain Magnetic Resonance Imaging (MRI) plays a crucial role in identifying the boundaries of abnormal regions associated with brain tumors, thereby facilitating more precise diagnosis and treatment planning. Previous studies have proposed automatic segmentation by leveraging deep learning. Convolution-based architectures are effective at spatial localization but are limited in capturing global contextual information, whereas transformer-based architectures can model long-range dependencies but often require substantial computational resources and may struggle to preserve fine-grained spatial details. To address these challenges, this research proposes Swin-Conv, a hybrid U-Net-based architecture consisting of a Swin Transformer encoder to capture the global context of an image and a convolutional decoder to preserve spatial localization during image reconstruction. Furthermore, the effectiveness of standard convolution and Mobile Inverted Bottleneck Convolution (MBConv) employed in the decoder is investigated across four Swin Transformer variants (Tiny, Small, Base, and Large). The experimental results on the public Low-Grade Glioma (LGG) MRI brain segmentation dataset demonstrate that the best performance is obtained by the Swin-Conv model with a standard convolutional decoder and the Swin-S. Comparative experiments with baseline models indicate that Swin-Conv achieves competitive performance with reasonable computational complexity. These findings highlight that Swin-Conv effectively integrates the benefits of a Swin Transformer encoder and a convolutional decoder to generate precise brain image segmentation efficiently, making it suitable for applied medical scenarios.

Keywords-brain tumor; deep learning; image segmentation; Swin Transformer; convolutional network

I. INTRODUCTION

The brain is a central organ that serves as the primary control center of the human body. Uncontrolled cell growth within or around the brain can lead to the formation of brain tumors, which may subsequently impair brain function and disrupt the central nervous system [1]. Global Cancer Statistics reported 321,731 new cases and 245,800 deaths from brain tumors worldwide in 2020, positioning brain tumors as the twelfth leading cause of cancer-related mortality [2]. The early diagnosis of brain tumors plays a crucial role in preparing suitable treatment planning and improving patients' survival rate [3].

In this context, advancements in medical imaging technologies, particularly Magnetic Resonance Imaging (MRI), have substantially enhanced the clinical assessment of brain tumors [4]. Accurate segmentation of brain MRI is crucial to delineate the precise boundaries of abnormal regions associated with brain tumors from other tissues, thereby facilitating more precise diagnosis and treatment planning [5]. However, brain tumor regions exhibit significant variations in shape, size, and location, and often present ambiguous boundaries with low contrast relative to surrounding healthy tissues. These characteristics make brain tumor segmentation a highly challenging task, particularly in accurately delineating tumor boundaries [6]. Manual segmentation of medical images performed by medical experts can be highly accurate, but it is typically time-consuming and costly. Therefore, the implementation of automatic image segmentation is essential to provide reliable and replicable results [7].

The convolutional encoder–decoder architecture is one of the most widely recognized deep learning models for image segmentation [8, 9]. U-Net, a specific type of encoder-decoder architecture, was introduced by authors in [10] for biomedical image segmentation. Recent studies have demonstrated the remarkable success of U-Net and its variants in brain tumor segmentation tasks. For instance, the original U-Net architecture has been reported to outperform SegNet and DeepLabv3 [11]. Furthermore, ResUNet was proposed by replacing the standard convolutional blocks with residual modules to facilitate more effective feature learning [12]. In another study, a U-Net-based architecture employing a pretrained DenseNet121 as the encoder achieved superior performance compared to the standard U-Net and ResNet-based models [13].

Several studies have further enhanced U-Net architectures by incorporating attention mechanisms to focus feature extraction on more relevant regions. For instance, Attention Res-UNet utilized attention gates within the Res-UNet backbone with a guided decoder [14]. Similarly, authors in [15] added an attention module in the decoder part, whereas TransUNetB incorporates a Transformer in its bottleneck layer [16]. Although variants of U-Net and the incorporated attention module have demonstrated substantial progress in medical image segmentation, the convolutional operation commonly employed in these architectures still exhibits inherent limitations in capturing long-range dependencies and global contextual information in an image, particularly in complex medical images such as brain MRI. Consequently, this limitation can lead to suboptimal segmentation performance when tumor boundaries exhibit high variability in shape, size, and location.

This limitation has motivated recent studies toward Transformer-based segmentation models. A Transformer network module consists only of a self-attention mechanism, which is effective in identifying long-range dependencies in data [17]. The Transformer architecture, known for its success in natural language processing, can be adapted for vision tasks by splitting images into patches. These patches are treated as tokens in the text domain. Vision Transformer (ViT), a modified Transformer network, especially designated for vision tasks, proved that a transformer could replace Convolutional Neural Networks (CNNs) in the classification and prediction of image patch sequences [18]. A pure Transformer-based architecture, referred to as the SEgmentation TRansformer (SETR), has demonstrated competitive results on several public segmentation datasets [19].

Despite the effectiveness of Transformer architectures in capturing long-range dependencies within data, pure Transformer architectures often require high computational costs and may struggle to preserve fine-grained spatial details. Therefore, recent studies have explored hybrid architectures that aim to combine the strengths of convolutional networks and Transformers. For instance, VcaNet integrates enhanced convolution, ViT, and a spatial attention module in the encoder, bottleneck, and decoder, respectively [20]. Similarly, UNet-VT integrates ViT as the encoder and UNet as the decoder [21]. Dual Vision Transformer-DSUNET integrates dual ViT in the encoder part with feature fusion and a convolutional network in the decoder [22].

However, the ViT models adopted in these studies rely on fixed-size image patches in their Multi-Head Self-Attention (MSA) mechanism, which may limit their ability to effectively capture visual elements with high spatial variability in images. Swin Transformer addresses this limitation by introducing two approaches: (1) a hierarchical feature representation that employs progressively larger patch sizes across successive layers, and (2) a shifted window approach for computing the MSA. The first approach is designed to handle variations in the scale of visual elements, whereas the second approach aims to improve computational efficiency and enables cross-window connections [23].

Motivated by the respective limitations of convolutional operations and pure Transformer architectures, this study proposes Swin-Conv, a hybrid U-Net-based architecture consisting of a Swin Transformer encoder and a convolutional decoder, specifically for brain tumor segmentation. The Swin Transformer enables effective and efficient modeling of long-range dependencies through hierarchical feature encoding and a shifted-window mechanism for MSA, whereas the convolutional decoder preserves precise spatial localization during image reconstruction. The proposed model demonstrates competitive performance compared with baseline models while maintaining computational efficiency. Furthermore, this study investigates the impact of different convolutional operations

employed in the decoder, as well as various Swin Transformer variants utilized in the encoder.

II. MATERIALS AND METHODS

This study proposes a hybrid architecture that utilizes the Swin Transformer and a convolutional network for medical image segmentation, specifically for brain MRI images. The overall research workflow and a detailed description of the proposed architectural model are explained in this section.

A. Data Collection and Division

This research utilizes the public Low-Grade Glioma (LGG) MRI brain segmentation dataset. This dataset was collected from The Cancer Imaging Archive (TCIA), is publicly available at [24], and was originally reported in [25]. This dataset consists of a collection of 3,929 brain slice MRI images with the corresponding manually segmented masks, comprising 2,556 images of normal cases and 1,373 images of tumor cases. Each image has a resolution of 256 × 256 pixels. The available dataset was divided into three independent subsets following a commonly adopted data split scenario in previous studies. Specifically, 2,750 images were allocated for the training subset, 786 images for the validation subset, and 393 images for the testing subset [26].

B. Data Preprocessing and Augmentation

Image preprocessing is essential to ensure that the image is ready and meets the requirements of the segmentation model. In this study, all images (both inputs and masks) were resized to 224 × 224 pixels to match the input dimension required by the pretrained Swin Transformer architecture on ImageNet. This approach is commonly adopted by research that applied pre-trained models [27], thus enabling effective use of pretrained weights while maintaining segmentation accuracy due to the relatively small difference from the original image size, especially given the limited size of the dataset. Subsequently, the input image was normalized to the range [0,1] and standardized to align the pixel value distribution with that used during the Swin Transformer pretraining on the ImageNet dataset. Each image pixel was subtracted by the mean and divided by the standard deviation of each corresponding RGB channel, using values of [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively. Meanwhile, the mask image was normalized into the range [0,1] using min–max normalization and then thresholded to 0 (background) or 1 (foreground), as the output of the final segmented image is a binary image.

Basic image augmentation techniques were applied to increase the variation of the image dataset at each epoch during model training. This augmentation plays a crucial role in ensuring that the model generalizes the patterns of the input image rather than memorizing them, thereby reducing the risk of overfitting [28]. The augmentation was applied only to the training images and their corresponding masks (ground truth). The image augmentation strategies were selected carefully to ensure that the augmented data remained representative of real medical image conditions as suggested in [26], including rotation with a maximum angle of 20° and horizontal flipping, each with a probability of 0.2.

C. Proposed Model Architecture

Following the popular U-Net architecture [10], the proposed Swin-Conv model consists of three parts, namely the encoder, bottleneck, and decoder, as illustrated in Figure 1. The encoder part utilizes a Swin Transformer [23] to extract the feature map. The bottleneck layer employs double convolutional operations to extract important feature maps before the reconstruction stage. Finally, the decoder part utilizes a convolutional network with transposed convolution for upsampling, followed by a convolution operation to reconstruct the output segmentation image. When the input and output dimensions match, a skip connection is utilized to retain the input signal and enhance gradient flow during training [10].

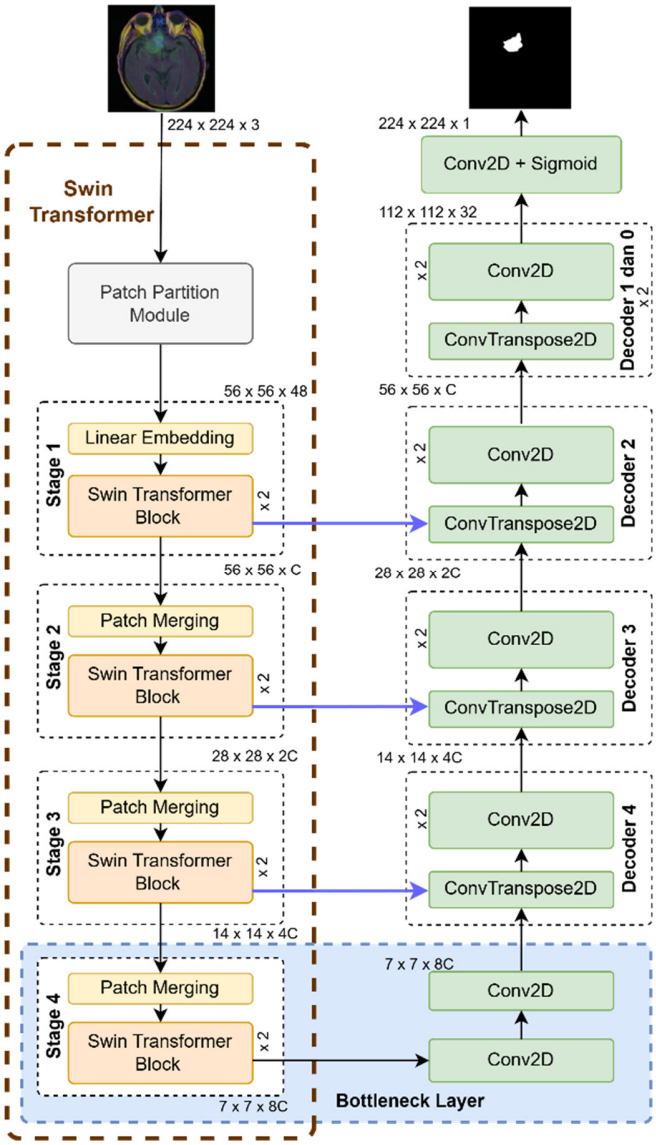

Fig. 1. The Swin-Conv model architecture.

1) Swin Transformer Encoder

The Swin Transformer is a variant of the Transformer architecture that adopts hierarchical feature representations combined with a shifted window approach for computing MSA. The input image for the Swin Transformer is of size 224 × 224 × 3. The Swin Transformer architecture begins with a patch partition module, followed by four consecutive stages, as illustrated in Figure 1. The input image is partitioned into non-overlapping patches of size 4 × 4 by the patch partition module, resulting in a spatial resolution of 56 × 56, with each patch having a dimension of 4 × 4 × 3 = 48. In Stage 1, a linear embedding projects each patch into an embedding dimension of C, followed by Swin Transformer blocks [23]. Each subsequent stage (Stage 2–4) comprises a patch merging layer followed by a Swin Transformer block. For example, in Stage 2, 2 × 2 adjacent patches from Stage 1 are merged by the patch merging layer, decreasing the spatial dimension to 28 × 28, while increasing the patch dimension to $4C$. Subsequently, the patch dimension is reduced to $2C$ through a linear layer transformation. The Swin Transformer block in each stage extracts features while maintaining the resolution. This process of patch merging and feature extraction is progressively repeated, producing spatial resolutions of 14 × 14 with a patch size of $4C$ in Stage 3, and 7 × 7 with a patch size of $8C$ in Stage 4. The use of different patch sizes at each stage aims to handle variations in the scale of visual elements. This approach also allows the model to construct hierarchical feature representations across all stages, similar to the feature representation obtained by typical convolutional networks architectures, such as VGG, ResNet, and EfficientNet [23].

A Swin Transformer block is composed of two successive components employing a Window-based MSA (W-MSA) and a Shifted-Window-based MSA (SW-MSA). The MSA computation is performed within each local window rather than across the entire image, which greatly improves computational efficiency. To enable cross-window connections, consecutive Swin Transformer blocks use different window configurations by shifting the partitioning such that each window overlaps with neighboring patches from the previous block [23].

There are four variants of the Swin Transformer architecture, namely Swin-T (Tiny), Swin-S (Small), Swin-B (Base), and Swin-L (Large). Each architecture differs in terms of feature embedding dimensions and the number of layers in each stage, which consequently leads to different levels of computational complexity [23]. Swin-T, Swin-S, Swin-B, and Swin-L employ embedding dimensions of 96, 96, 128, and 192, respectively. Among these variants, Swin-T utilizes the smallest number of layers, arranged as {2, 2, 6, 2}, whereas the other variants adopt a deeper configuration of {2, 2, 18, 2} [23].

Given the segmentation input image $\mathbf{X}$, the Swin Transformer encoder produces four hierarchical encoded feature representations from Stage 1 to Stage 4, denoted as $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, and $\mathbf{E}_4$, respectively. These hierarchical features are obtained using different patch dimensions, which are intended to effectively handle variations in the scale of visual elements [23], addressing the common challenge of multi-scale feature representation in medical image segmentation.

2) Bottleneck Layer

The bottleneck configuration follows the architectural principles of the original U-Net [10], ensuring a smooth transition between global feature extraction and local pixel-wise reconstruction. The bottleneck layer utilizes the output from the final stage of the encoder and applies two consecutive convolution operations with a kernel size of 3 × 3 [10]. This operation does not alter the spatial dimensions of the feature maps but aims to further refine and extract higher-level abstract features before the decoding stage.

The output of the bottleneck layer $\mathbf{B}$ can be obtained as:

$$\mathbf{B} = Conv_{3\times3}\big(Conv_{3\times3}(\mathbf{E}_4)\big) \quad (1)$$

where $Conv_{3\times3}$ represents a 2D convolution operation, the subscript denotes the kernel size.

3) Convolutional Network Decoder

A convolutional decoder is utilized to restore high-resolution spatial details that are often lost during the encoding process [10]. The decoder part employs transposed convolution operations to upsample the feature maps, thereby doubling their spatial dimensions in both width and height. A skip connection, indicated by the blue lines in Figure 1, is applied to concatenate the upsampled feature maps with the corresponding encoder feature maps of the same spatial size. This skip connection is used to preserve spatial details and fine-grained information that may have been lost during the downsampling process. Subsequently, the concatenated feature maps are passed through two consecutive convolutional layers with a 3 × 3 kernel to reduce the number of channels by half. The upsampling and convolution operations are performed iteratively five times until the feature map reaches the original input image dimension.

Let $\mathbf{D}_i$ denote the output of the decoder at level i. The output $\mathbf{D}_4$ is obtained from the bottleneck layer output ($\mathbf{B}$) and encoder output in Stage 3 ($\mathbf{E}_3$) as in (2). Subsequently, the output $\mathbf{D}_i$ (for $i = 3, 2$) is calculated from the previous decoder output $\mathbf{D}_{i+1}$ and the encoder output $\mathbf{E}_i$ by using (3). $ConvT_{2\times2}$ represents a 2D transposed convolution operation, where the subscript indicates the kernel size, and $[\cdot,\cdot]$ represents the concatenation operation used to implement the skip connection. The output $\mathbf{D}_1$ can be obtained from $\mathbf{D}_2$ without a skip connection, as shown in (4). The spatial resolution of $\mathbf{D}_1$ is half of the input image spatial size. Therefore, an additional decoder $\mathbf{D}_0$ is required to upsample the feature map back to the original input resolution, as shown in (5). Finally, a 1 × 1 convolution layer with a binary sigmoid activation function is applied to generate the segmentation output, as denoted in (6). Consequently, the network produces a segmentation map with a spatial size of 224 × 224 × 1.

$$\mathbf{D}_4 = Conv_{3\times3}\big(Conv_{3\times3}([ConvT_{2\times2}(\mathbf{B}), \mathbf{E}_3])\big) \quad (2)$$

$$\mathbf{D}_i = Conv_{3\times3}\big(Conv_{3\times3}([ConvT_{2\times2}(\mathbf{D}_{i+1}), \mathbf{E}_{i-1}])\big) \quad (3)$$

$$\mathbf{D}_1 = Conv_{3\times3}\Big(Conv_{3\times3}\big(ConvT_{2\times2}(\mathbf{D}_2)\big)\Big) \quad (4)$$

$$\mathbf{D}_0 = Conv_{3\times3}\left(Conv_{3\times3}\left(ConvT_{2\times2}(\mathbf{D}_1)\right)\right) \quad (5)$$

$$\mathbf{Y} = Sigmoid\left(Conv_{1\times1}(\mathbf{D}_0)\right) \quad (6)$$

4) Standard Convolution vs. Mobile Inverted Bottleneck Convolution

This research employs two types of convolution operations, namely standard convolution and Mobile Inverted Bottleneck Convolution (MBConv). Standard convolution is capable of capturing richer and more detailed feature representations, yet it requires a high computational cost and a large number of parameters [29]. In contrast, MBConv, originally introduced in MobileNet [30] and subsequently adopted in EfficientNet [31], significantly reduces both the number of parameters and the computational cost compared to standard convolution operations by applying pointwise and depthwise convolution. The Squeeze-and-Excitation (SE) block [32], originally incorporated in the MBConv layers of EfficientNet, was excluded from the proposed architecture to avoid redundancy with the Swin Transformer encoder's existing self-attention mechanisms. Its exclusion also prevents the loss of fine-grained spatial localization caused by global average pooling, which is critical for precise tumor boundary reconstruction. Furthermore, initial experimental results indicated that segmentation performance decreased when the SE mechanism was included.

D. Training and Validation

Two-dimensional medical image segmentation can be viewed as a binary classification task for each pixel. A classification task typically involves several stages, including training, validation, and testing. During training, the model is trained using a training dataset, and its performance is subsequently evaluated using a validation dataset under specific hyperparameter configurations. The training and validation processes are iteratively repeated with varying hyperparameter settings to identify the optimal model. Finally, the performance of the optimal model is assessed using a separate testing dataset.

E. Evaluation

Evaluation aims to assess the performance of the final model using the testing subset. The Dice coefficient and Intersection over Union (IoU) are employed as evaluation metrics in this study. The Dice coefficient measures the overlap or similarity between the predicted segmentation mask and the ground truth mask. The value ranges from 0 to 1, where 1 represents perfect overlap and 0 represents no overlap. The IoU is another segmentation metric that calculates the ratio of the intersection to the union of the predicted segmentation mask and the ground truth mask [9].

III. EXPERIMENTS AND RESULTS

This section presents and analyzes the performance of the proposed method for the brain MRI image segmentation task, specifically utilizing the LGG brain segmentation dataset. The proposed method follows a U-Net-like architecture, which employs the Swin Transformer as the encoder and a convolutional network as the decoder.

Based on variations in the convolutional operations within the decoder, the experiments in this study were organized into two main scenarios. The first scenario employed standard convolution operations in the decoder (referred to as Swin-Conv), whereas the second scenario utilized MBConv operations in the decoder (referred to as Swin-MBConv). Furthermore, each scenario was evaluated across four Swin Transformer variants: Tiny, Small, Base, and Large. All experiments were conducted on Google Colab Pro using an NVIDIA Tesla T4 GPU with approximately 16 GB of VRAM and 12 GB of system RAM.

Model training was conducted using a batch size of 32 for up to 150 epochs [26], with early stopping (patience = 10) applied to prevent overfitting. Epoch values beyond 150 were not further investigated, as all training scenarios consistently converged and stopped early before reaching the maximum number of epochs under this configuration. The Adam optimizer was employed for model optimization with a learning rate of 10^{-4} and a weight decay of 10^{-5}. Adam is widely adopted due to its adaptive learning rate adjustment, which enables faster and more stable convergence during training. Binary cross-entropy was selected as the loss function since the task is a binary segmentation [33].

A. Results on Training and Validation

Table I presents the experimental results of the first scenario (Swin-Conv) across four Swin Transformer variants during the training and validation phases. Overall, all Swin Transformer variants exhibited excellent segmentation performance, with Dice coefficients consistently exceeding 0.90 and IoU scores reaching or approaching 0.90 in both training and validation. While Swin-Conv-T slightly outperformed Swin-Conv-S during training, Swin-Conv-S yielded superior validation metrics. This suggests that increasing architectural depth facilitates a higher generalization ceiling by enabling the model to be more precise in learning complex features.

In contrast, Swin-Conv-B (utilizing a wider feature dimension of $C = 128$ with the same depth as Swin-Conv-S) recorded the lowest training performance, yet managed to surpass Swin-Conv-T during validation, though it still remained below Swin-Conv-S and Swin-Conv-L. This indicates that while the Swin-Conv-B possesses a higher capacity than the Swin-Conv-T, it appears to reside in a transitional complexity stage—large enough to complicate the optimization landscape, yet lacking the massive representational power required to surpass the more efficient Swin-Conv-S or the robust Swin-Conv-L. Meanwhile, Swin-Conv-L, utilizing the largest feature dimension ($C = 192$), achieved the highest performance in both training and validation. Collectively, these results imply that in this case, increasing architectural depth is a more critical determinant of model efficacy than merely expanding the feature width.

The experimental results of the second scenario (Swin-MBConv) across the four Swin Transformer variants during the training and validation phases are summarized in Table II. The experimental results generally indicate that the Swin-MBConv models achieve consistently high Dice coefficient and IoU

across all Swin Transformer variants. Swin-MBConv-T attained slightly higher training performance than Swin-MBConv-S; in contrast, the validation results show that Swin-MBConv-S exhibits better generalization capability, as reflected in its higher Dice coefficient and IoU than those of Swin-MBConv-T. Increasing model complexity further with the Swin-MBConv-B model resulted in the highest overall performance on the validation set, surpassing all other variants. Furthermore, although Swin-MBConv-L achieved the highest average Dice coefficient and IoU during the training phase, it recorded the lowest Dice coefficient and IoU during validation. This discrepancy indicates overfitting when employing a highly complex architecture. These findings suggest a trade-off between model complexity and segmentation performance. While a moderate increase in complexity from Swin-MBConv-T up to Swin-MBConv-B led to performance improvements, further increasing model complexity, as in Swin-MBConv-L, resulted in performance degradation rather than improvement.

TABLE I. TRAINING AND VALIDATION RESULTS OF THE SWIN-CONV MODEL

Model	Training Dice	Training IoU	Validation Dice	Validation IoU
Swin-Conv-T	0.9695	0.9490	0.9279	0.8999
Swin-Conv-S	0.9652	0.9416	0.9307	0.9011
Swin-Conv-B	0.9591	0.9336	0.9293	0.8983
Swin-Conv-L	0.9747	0.9556	0.9338	0.9056

TABLE II. TRAINING AND VALIDATION RESULTS OF THE SWIN-MBCONV MODEL

Model	Training Dice	Training IoU	Validation Dice	Validation IoU
Swin-MBConv-T	0.9408	0.9118	0.9182	0.8895
Swin-MBConv-S	0.9398	0.9103	0.9203	0.8909
Swin-MBConv-B	0.9498	0.9230	0.9254	0.8966
Swin-MBConv-L	0.9506	0.9228	0.9156	0.8863

The experimental results in Table I and II also reveal that Swin-Conv consistently achieves superior performance compared to Swin-MBConv in terms of Dice coefficient and IoU across all Swin Transformer variants in both the training and validation phases.

B. Results on Testing

Table III presents the testing results of the Swin-Conv architecture across four Swin Transformer variants. In the final test, Swin-Conv-T recorded the lowest performance, whereas Swin-Conv-S achieved the best performance with a Dice coefficient of 0.9350 and an IoU of 0.9076, followed by Swin-Conv-L and Swin-Conv-B. These results reveal that employing a shallower architecture is insufficient to achieve high generalization ability for unseen data.

Although Swin-Conv-L performed best in training and validation, Swin-Conv-S demonstrated more stability and maintained higher performance in testing. This may indicate that the Swin-Conv-S has a better balance between representation capacity and generalization ability on this dataset. In contrast, Swin-Conv-B, which was able to surpass the Swin-Conv-T, still remained below the Swin-Conv-S and Swin-Conv-L in testing. These findings strengthen the conclusion that increasing architectural depth is a more critical determinant of model efficacy than merely expanding the feature width in this brain tumor segmentation task.

The performance of Swin-MBConv on the testing data, as shown in Table IV, exhibited a similar pattern to that observed on the validation data. The lightest architecture, Swin-MBConv-T, still recorded lower performance compared with the larger variants, namely Swin-MBConv-S and Swin-MBConv-B. Swin-MBConv-S attained competitive results with Swin-MBConv-B. Among all variants, Swin-MBConv-S achieved the highest Dice coefficient of 0.9244 and IoU score of 0.8952. However, increasing feature width, as seen in the Swin-MBConv-L model, resulted in decreased performance despite achieving the highest accuracy during the training phase. This condition indicates overfitting, as the excessive capacity of the larger model led to over-parameterization relative to the available dataset. Consequently, the model tended to memorize the training data, thereby degrading its generalization ability on unseen data. Overall, these results suggest that increasing architectural depth from Swin-MBConv-T to Swin-MBConv-S effectively enhances performance; conversely, increasing the feature width or further scaling from Swin-MBConv-S to Swin-MBConv-B and Swin-MBConv-L leads to performance degradation.

TABLE III. TESTING RESULTS OF THE SWIN-CONV MODEL

Model	Testing Dice	Testing IoU
Swin-Conv-T	0.9237	0.8971
Swin-Conv-S	0.9350	0.9076
Swin-Conv-B	0.9335	0.9065
Swin-Conv-L	0.9337	0.9070

TABLE IV. TESTING RESULTS OF THE SWIN-MBCONV MODEL

Model	Testing Dice	Testing IoU
Swin-MBConv-T	0.9107	0.8813
Swin-MBConv-S	0.9244	0.8952
Swin-MBConv-B	0.9232	0.8948
Swin-MBConv-L	0.9042	0.8733

The overall comparison of the average Dice coefficient and IoU between Swin-Conv and Swin-MBConv on the testing set, as shown in Figure 2, also exhibits a pattern similar to that observed on the validation set. Specifically, Swin-Conv consistently outperforms Swin-MBConv across all Swin Transformer variants. This finding suggests that, for medical image segmentation tasks requiring detailed structural reconstruction, standard convolution is more appropriate than MBConv, as standard convolution allows the model to preserve and recover fine-grained spatial details while reconstructing high-quality segmentation results. In contrast, the depthwise separable convolution in MBConv may limit the capability to capture highly complex spatial–channel dependencies that are essential for precise image reconstruction.

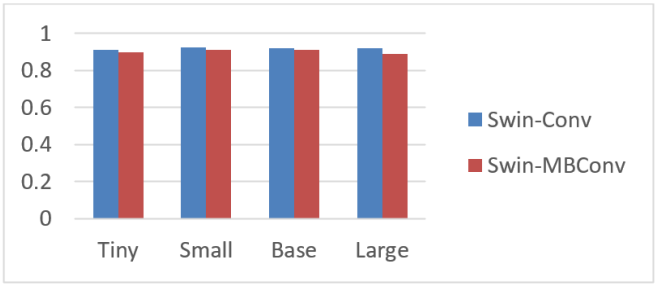

Fig. 2. Average Dice coefficient and IoU of Swin-Conv and Swin-MBConv.

Figure 3 presents several sample cases consisting of the input image, its corresponding ground truth mask, the segmentation mask generated by Swin-Conv, and the segmentation mask generated by Swin-MBConv, respectively. These results demonstrate the effectiveness of Swin-Conv in delineating object boundaries and producing more precise and accurate segmentation outputs compared to Swin-MBConv.

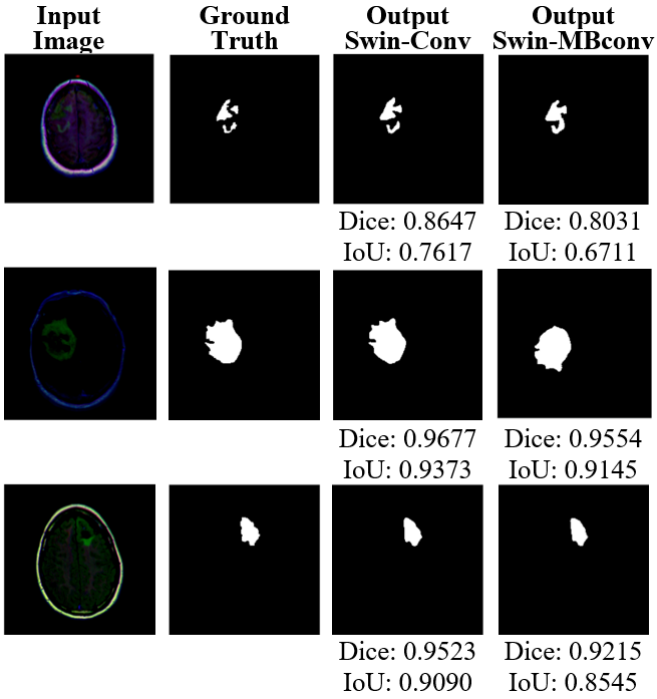

Fig. 3. Ground truth and mask output from Swin-Conv and Swin-MBConv.

C. Comparison with Baseline Models and Previous Studies

This study also compares the performance, complexity, and inference time of the proposed method with the baseline methods, namely U-Net (pure convolutional encoder–decoder) and Swin-UNet (pure transformer encoder–decoder), specifically on the LGG MRI segmentation dataset, as depicted in Table V.

It can be observed that the proposed Swin-Conv model demonstrates competitive segmentation performance by achieving a higher mean Dice coefficient than the baseline methods and a lower standard deviation, indicating more consistent segmentation performance across different test images. Furthermore, a statistical significance test performed on the Dice coefficient and IoU of the testing samples using the Wilcoxon signed-rank test yielded p -values < 0.05 for comparisons with U-Net as well as Swin-UNet, indicating that the performance improvement achieved by the proposed Swin-Conv model is statistically significant at the 5% level.

TABLE V. PERFORMANCE COMPARISON OF SWIN-CONV AND BASELINE MODELS

Model	Testing Dice		Testing IoU	
	Mean ± Stdev	*p*-value	Mean ± Stdev	*p*-value
Swin-Conv	0.9350 ± 0.1698	-	0.9076 ± 0.1913	-
U-Net	0.9222 ± 0.2033	0.0361	0.8963 ± 0.2200	0.0333
Swin-UNet	0.9298 ± 0.1830	0.0058	0.9028 ± 0.2043	0.0046

In addition, to evaluate the robustness of the proposed model, model training was repeated three times using different random seed values in each run, specifically 42, 123, and 2023. Subsequently, the resulting model was evaluated on the same testing dataset. The average Dice coefficient and IoU from these three evaluations are 0.9372 ± 0.0036 and 0.9097 ± 0.0043, respectively. The low standard deviation indicates the robustness of the proposed model across repeated training.

The proposed model also achieves a favorable balance between model complexity (number of parameters and FLOPs) and efficiency (inference time), as presented in Table VI. Although the U-Net model has the lowest number of parameters, it requires the highest FLOPs, indicating a larger number of arithmetic operations.

However, since these operations can be efficiently parallelized on GPUs, U-Net achieves the shortest inference time. In contrast, Swin-UNet has the lowest FLOPs, but its operations are less GPU-efficient, resulting in a longer inference time. The reasonable balance between model complexity and inference time makes the proposed model suitable for applied medical scenarios.

TABLE VI. COMPUTATIONAL COMPLEXITY COMPARISON OF SWIN-CONV AND BASELINE MODELS

Model	#Parameters	FLOP (GFLOPs)	Inference time (ms)
Swin-Conv	66,355,115	14.67	31.58
U-Net	31,043,521	41.91	19.29
Swin-UNet	69,743,467	13.11	41.59

Finally, the comparison of the Swin-Conv with the reported results from previous studies is presented in Table VII. It can be seen that Swin-Conv demonstrates improved performance over other proposed segmentation methods from previous studies. The proposed Swin-Conv model leverages the strengths of the Swin Transformer in capturing global contextual dependencies during encoding and convolutional networks in preserving precise spatial localization during decoding, thereby resulting in improved segmentation performance.

TABLE VII. PERFORMANCE COMPARISON OF SWIN-CONV AND OTHER MODELS FROM PREVIOUS STUDIES

Model	Testing Dice	Testing IoU
Swin-Conv (proposed method)	0.9350	0.9076
U-Net-based architecture with attention mechanisms and Bayesian Optimization [15]	0.8969	0.8151
Novel ResUNet [12]	0.9056	0.8293
ECASE-UNet (efficient channel attention and SE) [26]	0.9197	0.8521

IV. CONCLUSION

This study proposes a hybrid segmentation model that follows a U-Net-like architecture, which employs the Swin Transformer as the encoder and a convolutional network as the decoder. Furthermore, this study investigates the effectiveness of two convolutional operations in the decoder, namely standard convolution in the Swin-Conv model and Mobile Inverted Bottleneck Convolution (MBConv) in the Swin-MBConv model, across four Swin Transformer variants: Tiny, Small, Base, and Large. The experimental results on the public Low-Grade Glioma (LGG) MRI brain segmentation dataset demonstrate that Swin-Conv consistently achieves superior performance compared to Swin-MBConv, highlighting the suitability of standard convolution operations for medical image segmentation tasks that require detailed structural reconstruction.

The results across different Swin Transformer variants further reveal that increasing architectural depth is a more critical determinant of model efficacy than expanding the feature width. In this study, Swin-Conv utilizing a standard convolutional decoder and the Swin-S (Small) variant achieves the highest performance. Comparative experiments with baseline models (U-Net and Swin-UNet) indicate that Swin-Conv achieves competitive segmentation performance while maintaining reasonable computational complexity.

These findings highlight that the proposed Swin-Conv model effectively integrates the benefits of a Swin Transformer encoder and a convolutional decoder to generate precise brain image segmentation efficiently, making it suitable for applied medical imaging scenarios.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the Faculty of Science and Mathematics, Universitas Diponegoro, under the Riset Utama A research grant scheme with the contract number of 24.B/UN7.F8/PP/II/2025.

DATA AVAILABILITY

The LGG MRI brain segmentation dataset used in this study is publicly available at [24] and was originally reported in [25].

REFERENCES

[1] K. C. Pasunoori, Ch. R. Prasad, and K. R. Kumar, "A systematic review on deep learning based brain tumor segmentation and detection using MRI: Past insights, present techniques and future trends," *Computational Biology and Chemistry*, vol. 120, Feb. 2026, Art. no. 108696, https://doi.org/10.1016/j.compbiolchem.2025.108696.

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, https://doi.org/10.3322/caac.21660.

[3] M. Pichaivel, G. Anbumani, P. Theivendren, and M. Gopal, "An Overview of Brain Tumor," in *Brain Tumors*, A. Agrawal, Ed. London, UK: IntechOpen, 2022, https://doi.org/10.5772/intechopen.100806.

[4] M. Martucci *et al.*, "Magnetic Resonance Imaging of Primary Adult Brain Tumors: State of the Art and Future Perspectives," *Biomedicines*, vol. 11, no. 2, Jan. 2023, Art. no. 364, https://doi.org/10.3390/biomedicines11020364.

[5] Z. Yi, L. Long, Y. Zeng, and Z. Liu, "Current Advances and Challenges in Radiomics of Brain Tumors," *Frontiers in Oncology*, vol. 11, Oct. 2021, Art. no. 732196, https://doi.org/10.3389/fonc.2021.732196.

[6] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "Computer-Aided Grading of Gliomas Combining Automatic Segmentation and Radiomics," *International Journal of Biomedical Imaging*, vol. 2018, no. 1, May 2018, Art. no. 2512037, https://doi.org/10.1155/2018/2512037.

[7] A. A. Adegun, R. O. Ogundokun, M. O. Adebiyi, and E. O. Asani, "CAD-Based Machine Learning Project for Reducing Human-Factor-Related Errors in Medical Image Analysis," in *Handbook of Research on the Role of Human Factors in IT Project Management*, S. Misra and A. Adewumi, Eds. Hershey, PA, USA: IGI Global Scientific Publishing, 2020, pp. 164–172, https://doi.org/10.4018/978-1-7998-1279-1.ch011.

[8] Md. E. Rayed, S. M. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, Jan. 2024, Art. no. 101504, https://doi.org/10.1016/j.imu.2024.101504.

[9] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, July 2022, https://doi.org/10.1109/TPAMI.2021.3059968.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[11] N. C. Kundur, H. R. Divakar, S. Khaiyum, K. P. Rakshitha, P. M. Dhulavvagol, and A. S. Meti, "Deep Neural Networks for Precise Brain Tumor Delineation: A U-Net and TensorFlow Approach," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23686–23691, June 2025, https://doi.org/10.48084/etasr.10684.

[12] P. Santosh Kumar, V. P. Sakthivel, M. Raju, and P. D. Sathya, "Brain tumor segmentation of the FLAIR MRI images using novel ResUnet," *Biomedical Signal Processing and Control*, vol. 82, Apr. 2023, Art. no. 104586, https://doi.org/10.1016/j.bspc.2023.104586.

[13] N. Cinar, A. Ozcan, and M. Kaya, "A hybrid DenseNet121-UNet model for brain tumor segmentation from MR Images," *Biomedical Signal Processing and Control*, vol. 76, July 2022, Art. no. 103647, https://doi.org/10.1016/j.bspc.2022.103647.

[14] D. Maji, P. Sigedar, and M. Singh, "Attention Res-UNet with Guided Decoder for semantic segmentation of brain tumors," *Biomedical Signal Processing and Control*, vol. 71, Jan. 2022, Art. no. 103077, https://doi.org/10.1016/j.bspc.2021.103077.

[15] K. Ramalakshmi and L. Krishna Kumari, "U-Net-based architecture with attention mechanisms and Bayesian Optimization for brain tumor segmentation using MR images," *Computers in Biology and Medicine*, vol. 195, Sept. 2025, Art. no. 110677, https://doi.org/10.1016/j.compbiomed.2025.110677.

[16] K. G. Khushubu *et al.*, "TransUNetB: An advanced Transformer–UNet framework for efficient and explainable brain tumor segmentation," *Informatics in Medicine Unlocked*, vol. 59, Jan. 2025, Art. no. 101706, https://doi.org/10.1016/j.imu.2025.101706.

[17] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010, https://doi.org/10.48550/arXiv.1706.03762.

[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, Vienna, Austria, 2021, https://doi.org/10.48550/arXiv.2010.11929.

[19] S. Zheng *et al.*, "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 6877–6886, https://doi.org/10.1109/CVPR46437.2021.00681.

[20] D. Pan, J. Shen, Z. Al-Huda, and M. A. A. Al-qaness, "VcaNet: Vision Transformer with fusion channel and spatial attention module for 3D brain tumor segmentation," *Computers in Biology and Medicine*, vol. 186, Mar. 2025, Art. no. 109662, https://doi.org/10.1016/j.compbiomed.2025.109662.

[21] S. Kannan, S. M, V. Balaji, and R. P. Singh, "UNet-VT: Integrating U-Net and Vision Transformers for Enhancing Brain Tumor Segmentation in MRI scans," *Procedia Computer Science*, vol. 258, pp. 2210–2219, Jan. 2025, https://doi.org/10.1016/j.procs.2025.04.471.

[22] M. Zakariah, M. Al-Razgan, and T. Alfakih, "Dual vision Transformer-DSUNET with feature fusion for brain tumor segmentation," *Heliyon*, vol. 10, no. 18, Sept. 2024, Art. no. e37804, https://doi.org/10.1016/j.heliyon.2024.e37804.

[23] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, 2021, pp. 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986.

[24] "Brain MRI segmentation." Kaggle. [Online]. Available: https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation.

[25] M. A. Mazurowski, K. Clark, N. M. Czarnek, P. Shamsesfandabadi, K. B. Peters, and A. Saha, "Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data," *Journal of Neuro-Oncology*, vol. 133, no. 1, pp. 27–35, May 2017, https://doi.org/10.1007/s11060-017-2420-1.

[26] H. Hedibi, M. Beladgham, and A. Bouida, "A combined attention mechanism for brain tumor segmentation of lower-grade glioma in magnetic resonance images," *Computers in Biology and Medicine*, vol. 193, July 2025, Art. no. 110380, https://doi.org/10.1016/j.compbiomed.2025.110380.

[27] I. Aboussaleh, J. Riffi, K. E. Fazazy, M. A. Mahraz, and H. Tairi, "Efficient U-Net Architecture with Multiple Encoders and Attention Mechanism Decoders for Brain Tumor Segmentation," *Diagnostics*, vol. 13, no. 5, Feb. 2023, Art. no. 872, https://doi.org/10.3390/diagnostics13050872.

[28] P. A. Abdalla, B. A. Mohammed, and A. M. Saeed, "The impact of image augmentation techniques of MRI patients in deep transfer learning networks for brain tumor detection," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Nov. 2023, Art. no. 51, https://doi.org/10.1186/s43067-023-00119-9.

[29] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, Paris, France, 2010, pp. 253–256, https://doi.org/10.1109/ISCAS.2010.5537907.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.

[31] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, 2019, pp. 6105–6114, https://doi.org/10.48550/arXiv.1905.11946.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132–7141, https://doi.org/10.1109/CVPR.2018.00745.

[33] J. Krohn, G. Beyleveld, and A. Bassens, *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*, 1st ed. Boston, MA, USA: Addison-Wesley Professional, 2019.