Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Comparative Evaluation of CNN and Transformer-Based Models for Brain Tumor MRI Segmentation

Yakup Kuzhan[1] and Mücahid Günay[2]✉
[1]*Graduate School of Natural and Applied Sciences, Department of Information Systems, Kahramanmaraş Sütçü İmam University, Kahramanmaraş, Türkiye*
[2]*Department of Computer Engineering, Kahramanmaraş Sütçü İmam University, Kahramanmaraş, Türkiye*
Corresponding Author: Mücahid Günay, **E-mail**: mucahidgunay@gmail.com

| ABSTRACT
Brain tumors represent a critical medical problem with significant impact on human health. Manual assessment processes are insufficient in addressing this problem due to the difficulty in identifying tumor regions and disagreements among experts. This has led to an increasing demand for automated tumor detection or segmentation using machine learning and deep learning methods. This study compares the segmentation performance of selected deep learning models to investigate which model architecture yields better results under specific conditions. Two main approaches are considered: CNN-based (FCN, U-Net, Attention U-Net, DeepLabV3+) and a Transformer-based model (Swin U-Net). To ensure a fair comparison, all models were trained on the same dataset (BRISC 2025) under identical training conditions. Model performance was evaluated using Dice, IoU, Precision, Recall, and HD95 metrics. Among the tested models, DeepLabV3+ achieved the best performance with a Dice score of 0.8575 and an HD95 value of 7.8367. The findings indicate that CNN-based models outperform the Transformer-based model under the given dataset and experimental conditions. The results suggest that the performance of Transformer-based models may be sensitive to dataset characteristics. This study contributes to the literature by systematically evaluating CNN- and Transformer-based models under identical experimental settings.

| KEYWORDS

Brain Tumor Segmentation, U-Net, DeepLabV3+, Swin U-Net, BRISC 2025

| ARTICLE INFORMATION

ACCEPTED: 05 April 2026 **PUBLISHED:** 17 May 2026 **DOI:** 10.32996/jcsts.2026.8.7.5

1. Introduction

Tumor cells are characterized by uncontrolled proliferation in the body. Brain tumors refer to abnormal growths located in the brain or surrounding tissues. They are divided into different subgroups according to their location or cell type (Pichaivel et al., 2022). Brain tumors negatively affect an individual's quality of life. Furthermore, their negative impact on life expectancy is quite significant (Ostrom et al., 2019).

Early detection and diagnosis have a significant impact on an individual's quality of life and lifespan. However, manual tumor detection by specialist physicians leads to limitations such as the inability to precisely determine the boundaries of the cell and the emergence of differing opinions among physicians. These problems negatively affect the treatment process (Seshimo et al., 2024). Comprehensive reviews of the historical development and current status of deep learning methods in brain MRI segmentation clearly demonstrate the potential of this algorithm to eliminate inconsistencies in manual assessments (Akkus et al., 2017).

Globally, the increasing number of patients reveals that manual segmentation or tumor detection is not sustainable on a clinical scale. Consequently, there has been an increased trend towards semi-automatic and automatic segmentation methods. Machine learning and deep learning-based methods enable highly accurate and sustainable results by minimizing human intervention (Iqbal et al., 2023).

Deep learning-based methods enable the precise detection of detailed structural and textural differences that the human eye has difficulty distinguishing. These methods, designed specifically for a particular task, offer higher accuracy rates (Abidin et al., 2024). Numerous architectures have been developed within the deep learning framework. The most frequently used architectures are CNN-based architectures with very high automatic feature extraction capacity (Mostafa et al., 2023). In addition, Transformer-based architectures with strong global feature capture capabilities have recently gained attention in the literature (Liu et al., 2021).

Within CNN-based architectures, multi-scale contextual feature capture mechanisms, particularly ASPP (Atrous Spatial Pyramid Pooling), have yielded successful results in the literature (Chen et al., 2018).

The segmentation and classification performance of deep learning architectures has been examined in a limited number of studies conducted on the BRISC 2025 dataset. A benchmark study was conducted by Fateh et al. (2026). In this study, three different tumor types, glioma, meningioma, and pituitary tumors, were considered, and their segmentation performances were evaluated based on tumor type. CNN-based models such as U-Net, U-Net++, DeepLabV3+, LinkNet, and MANet, and Transformer-based architectures such as SaberNet and ABANet were used in the evaluation. The SaberNet model showed the highest performance with 80.6% weighted mIoU. In contrast, classical CNN approaches such as U-Net also provided a strong baseline with 75.7% weighted mIoU (Fateh et al., 2026). The weighted mIoU metric provides a more representative evaluation by considering class distribution across tumor types. In the reference study, the models were trained using standard Binary Cross-Entropy and Dice loss functions. However, this approach can limit the focusing ability of the models in cases where the tumor region in medical images is very small compared to the background.

Another study conducted on the same dataset is by Bhamboo et al. This study proposes an uncertainty-aware segmentation approach. A heterogeneous ensemble architecture including UNet2.5D, AttentionUNet2.5D, and TransUNet models was used, and epistemic uncertainty estimation was performed using the Monte Carlo Dropout method. In the ensemble model used, the images were at a resolution of 128x128, and an average Dice score of 0.798 was obtained. This study indicates that combining Transformer-based models with CNN-based models can increase segmentation reliability. At the same time, the preference for low-resolution input images such as 128x128 is highlighted as a limitation that makes it difficult to preserve the details and boundary sharpness of small tumor structures (Bhamboo et al., 2025).

This study aims to systematically evaluate the strengths and weaknesses of different model architectures under controlled experimental conditions. In medical image segmentation, the relatively small size of tumor regions compared to the total image reveals a significant class imbalance that complicates the learning process of the models.

To overcome the limitations of commonly used standard loss functions in managing class imbalance and to ensure that models focus on small tumor regions, a hybrid loss function combining Focal Tversky and Focal Dice was employed in this study. In addition, unlike studies using low-resolution inputs, the boundary sensitivities of the models were examined in more detail using input data with a resolution of 384 x 384 pixels. CNN- and Transformer-based models were compared on the BRISC 2025 dataset using consistent evaluation metrics to ensure a fair comparison. The model architecture that achieved the best performance was analyzed. This study provides a controlled comparison of CNN- and Transformer-based models using the same dataset and evaluation metrics to identify the most effective architecture.

2. Material and Methods

2.1 Dataset

The data used in this study were obtained from the BRISC 2025 dataset. This dataset is a recently introduced dataset with expert-annotated high-resolution (512 x 512) images, prepared for use in deep learning-based segmentation studies. The dataset contains a total of 4793 images for the segmentation task. These images were divided into 3933 training images and 860 test images following the original dataset split.

All images are derived from contrast-enhanced T1-weighted (CE-T1) MRI scans and are provided in 8-bit grayscale JPG format. One of the key characteristics of the dataset is that it includes images from three anatomical planes: axial, coronal, and sagittal. This leads to the same tumor structure appearing in significantly different morphological appearances and sizes across different slices, which poses a significant challenge for spatial generalization in deep learning models. (Fateh et al., 2026)

2.2 Preprocessing

To ensure fair and accurate evaluation of model performance, the dataset was divided into three clusters: training, validation, and testing. This split enables a reliable evaluation of model performance.

The training set is the primary data used for model learning process. The validation dataset was used to monitor the models' performance during the training process. The validation set was created using a fixed random seed (random seed=42) to obtain the same splitting results in different studies (Virtanen et al., 2020). The test dataset was used for final model evaluation after the model training was completed. The models had no access to the test data during training.

The dataset was split according to ratios commonly used in the literature. In the original folder structure of the dataset, the separation of training (3933) images and test (860 images) was predetermined. In this study, the test set was kept, while the

training set was split into 85% - 15% for model training and validation purposes. In this context, 85% of the training data was used in the model's learning process, while the remaining 15% is reserved as a validation dataset. This separation enables monitoring the models' performance during the training process in terms of overfitting. The test set consists of 860 images from the original folder structure, and this data was not included in any splitting process at any stage. This approach enables a more reliable evaluation of model performance.

The images in the dataset were rescaled to 384 x 384 pixels to reduce computational costs and to meet hardware constraints. Furthermore, segmentation masks were converted to binary format to match the model output format. Additionally, data augmentation methods were not applied in this study due to the sufficient size and natural variability of the dataset. The inclusion of different anatomical planes in the BRISC 2025 dataset allows models to learn from different spatial perspectives.

In medical images, the tumor area is very small compared to the total image area, which can limit the performance of standard loss functions. However, Dice is based on the overlap between predicted and ground truth pixels to resolve these class imbalances and offers a more balanced learning process. The standard Dice coefficient DSC and the corresponding Dice loss L_{Dice} are defined as follows.

$$\text{DSC}=\frac{2*\sum p_i\, g_i+\epsilon}{\sum p_i+\sum g_i+\epsilon} \tag{1}$$

Here, p_i denotes the predicted pixel probability, g_i denotes the ground truth label, and ϵ is a smoothing term to prevent division by zero.

The Focal Dice loss is an extension of the standard Dice loss designed to improve model performance in cases of class imbalance or hard-to-segment regions. The Focal Dice Loss L_{FD} is defined as follows:

$$L_{\text{FD}}=(1\text{-DSC})^{\gamma} \tag{2}$$

In the formula, DSC represents the Dice Similarity Coefficient. According to Lin et al. (2017), Focal Dice Loss is equal to standard Dice when $\gamma=0$. As γ increases, the model reduces the weight given to easy samples and increases its focus on difficult pixels. The focusing parameter γ was set to 1.3 in this study, considering the value of 2 suggested by Lin et al. (2017) and the range of 1-2 suggested by (Abraham & Khan, 2019).

The Tversky index, a generalized form of the Dice coefficient, enables the separate weighting of false positive (FP) and false negative (FN) errors through the α and β parameters. This provides flexibility with respect to class imbalance (Salehi et al., 2017). This flexibility makes it possible to penalize false negative (FN) tumor pixels, which are clinically critical, especially in medical images, more heavily. The Tversky Index TI used in Focal Tversky is formulated as follows:

$$\text{TI}=\frac{\sum p_i\, g_i+\epsilon}{\sum p_i\, g_i+\alpha*\sum p_i\,(1\text{-}g_i)+\beta*\sum(1\text{-}p_i)\, g_i+\epsilon} \tag{3}$$

In this study, the aim is to reduce the risk of the model missing tumor regions by assigning more importance to false negatives than false positives by assigning weights to $\alpha=0.7$ and $\beta=0.3$. The Focal Tversky Loss obtained by incorporating the focal component reduces the influence of easy examples and emphasizes difficult pixels (Abraham & Khan, 2019). Focal Tversky Loss L_{FT} is formulated as follows:

$$L_{\text{FT}}=(1\text{-TI})^{1/\gamma} \tag{4}$$

The hybrid function used in the study was constructed by combining the Focal Dice and Focal Tversky loss functions. The equal weighting (0.5-0.5) of both functions aims to optimize the Focal Tversky's focusing capability on difficult pixels and Focal Dice's overlap-based representation in a balanced way. The formulation is given as follows:

$$L_{\text{Combined}}=0.5*L_{\text{FT}}+0.5*L_{\text{FD}} \tag{5}$$

2.3 Model Architectures

FCN (Fully Convolutional Network) is based on redesigning image classification networks to create models capable of generating pixel-level predictions. Due to its full convolutional structure, features obtained from the input image are directly converted into pixel-wise segmentation output (Long et al., 2015). In the encoder stage, spatial resolution is reduced through sequential

convolution and pooling operations, resulting in more abstract feature representations. In this study, a four-stage encoder structure was employed, with two sequential convolutions, Batch Normalization, and LeakyReLU activation applied in each stage. The number of filters was progressively increased as 32, 64, 128, and 256, and the spatial dimension was halved at the end of each stage using MaxPooling. In the decoder stage, these features are transferred back to the input resolution via upsampling layers. Due to its full convolutional structure, the model generates class probabilities for each pixel in the input image, and the loss function is computed at the pixel level. Since this structure does not include skip connections between the encoder and decoder, capturing fine boundary details may be limited. This design was deliberately adopted to maintain architectural simplicity.

U-Net is a convolutional neural network model developed for medical image segmentation. Due to its encoder-decoder structure and skip connections, it achieves high segmentation accuracy by integrating spatial and semantic features (Ronneberger et al., 2015). In this study, an EfficientNetB0 backbone, pre-trained on ImageNet was used as the encoder. This choice leverages transfer learning advantages in medical imaging tasks, improving the model's feature extraction capability. In the decoder stage, feature maps are progressively restored to the input resolution via Conv2DTranspose layers. At each stage, the corresponding feature maps from the encoder are concatenated. The number of filters is progressively reduced as 256, 128, 64, and 32. Skip connections directly transfer all feature maps, which may introduce irrelevant background information into the decoder stage. This limitation motivates the development of attention mechanisms.

Attention U-Net is based on the encoder-decoder architecture of the standard U-Net model. Its key feature is the use of Attention Gates to filter features from skip connections, enabling the model to focus on relevant tumor regions while suppressing irrelevant background information (Schlemper et al., 2019). In this study, Attention U-Net was trained from scratch, and no pre-trained backbone was used. Following the original architecture, the number of filters in the encoder stage was progressively increased as 64, 128, 256, and 512, reaching 1024 filters in the bottleneck layer. In the decoder stage, upsampling is performed via Conv2DTranspose layers, and the corresponding encoder features are concatenated after passing through the Attention Gate mechanism in each stage. The Attention Gate structure combines the feature map from the encoder with the contextual signal from the decoder to generate an attention coefficient in the range [0, 1] for each pixel, and this coefficient is element-wise multiplied with the feature map to suppress irrelevant activations. Since this mechanism is based on the soft-attention principle, it is optimized via end-to-end backpropagation (Schlemper et al., 2019). This choice was adopted to evaluate the contribution of the attention mechanism independently of pre-trained backbone feature transfer. Rather than directly transferring all feature maps through skip connections, this approach filters them using attention gates, preventing irrelevant background information from being propagated to the decoder stage.

DeepLabV3+ is an encoder-decoder architecture designed to improve segmentation boundary accuracy while capturing multi-scale contextual information. The ASPP (Atrous Spatial Pyramid Pooling) module, integrated into the ResNet50 backbone in the encoder stage, captures features at multiple scales by applying parallel dilated convolutions with atrous rates of 6, 12, and 18. Atrous convolution expands the receptive field without increasing the number of parameters, enabling high-resolution feature extraction. In the decoder stage, the ASPP output is upsampled using bilinear interpolation, and low-level feature maps from early layers of the backbone are combined after channel reduction via 1x1 convolution. Subsequently, 3x3 convolution layers refine boundary details, and the final upsampling step produces pixel-wise segmentation output (Chen et al., 2018). In this study, the DeepLabV3+ architecture was implemented with a ResNet50 backbone pre-trained on ImageNet. Additionally, due to numerical instability caused by mixed precision, the model was trained using float32 precision.

Swin U-Net is an encoder-decoder architecture adapted from the Swin Transformer for semantic segmentation (Cao et al., 2021). Unlike traditional convolutional neural networks (CNNs), it captures both local and global contextual dependencies through a window-based self-attention mechanism (Liu et al., 2021). In the encoder stage, the input image is processed through hierarchical Swin Transformer blocks, which enable information flow across different regions via the shifted window mechanism. In the decoder stage, feature maps are progressively upsampled using patch expanding layers, and spatial information is preserved through skip connections between the encoder and decoder layers. In this study, Swin U-Net was trained from scratch. Despite the high number of parameters, the batch size was set to 10 for this model.

3. Training and Evaluation

The hyperparameters used during the training process are summarized in Table 1. The Adam optimization algorithm was employed to ensure stable convergence of the models (Kingma & Ba, 2014). The number of epochs was set at 50, and this value was selected to ensure sufficient convergence while keeping the risk of overfitting under control. A hybrid function combining Focal Dice and Focal Tversky was employed to improve segmentation performance, particularly for small tumor regions.

Table 1. Hyperparameters used during model training

Parameter	Value
Number of epochs	50
Early Stopping (patience)	12
Batch size	10
Learning rate	2 x 10-4
Optimizer	Adam
Loss function	Focal Dice + Focal Tversky (Hybrid)
Input size	384 × 384

The evaluation metrics are defined based on four fundamental concepts: True Positive (TP), pixels correctly predicted as tumor regions; False Positive (FP), pixels incorrectly predicted as tumor regions while belonging to normal tissue; True Negative (TN), pixels correctly predicted as normal tissue; False Negative (FN), pixels incorrectly predicted as normal tissue while belonging to tumor regions.

The metrics used for evaluation are Dice, IoU, Precision, Recall, and HD95. The Dice coefficient measures the overlap between the predicted mask and the ground truth mask, providing a robust measure of segmentation performance.

$$Dice = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{6}$$

Due to its small size and class imbalance of tumor regions, the Dice coefficient is widely used in medical image segmentation (Milletari et al., 2016). IoU measures the ratio of the intersection over the union between the predicted region and the ground truth region. Since IoU is based on both intersection and union, it provides a stricter evaluation against segmentation errors.

$$IoU = \frac{TP}{TP + FP + FN} \tag{7}$$

Precision measures the proportion of pixels predicted as tumor that are actually tumor, indicating the model's tendency to produce false positives.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall, on the other hand, measures the proportion of actual tumor pixels correctly identified by the model and is directly related to the clinically critical false negative rate.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

In applications where a low margin of error is critical, such as brain tumor segmentation, missing tumor regions can lead to serious clinical consequences, making this metric particularly important. HD95 is a metric that measures the distance between the predicted tumor boundaries and the ground truth tumor boundaries. Unlike the standard Hausdorff distance, it is based on the 95th percentile of the boundary distances, thereby reducing the effect of outlier boundary pixels. This allows for a more reliable assessment of boundary accuracy. The basic Hausdorff Distance (HD) formula is defined as follows:

$$HD(G, P) = max\left(d_H(G, P), d_H(P, G)\right) \tag{10}$$

In the formula, G (Ground Truth) and P (predicted set) are defined as two sets of points. The Hausdorff distance $d_H(G, P)$ is defined as the maximum distance from a point in one set to the nearest point in the other set. It can be interpreted as how far a point on the ground truth boundary can be from the predicted boundary. Mathematically, it is expressed as follows:

$$d_H(G, P) = max_{g \in G} min_{p \in P} \|g\text{-}p\| \tag{11}$$

Here, the expression $||g\text{-}p||$ represents the Euclidean distance between points in sets G and P. The formula calculates the shortest distances from each point in set G to set P and outputs the maximum of these values. This bidirectional calculation mechanism enables the evaluation of the spatial boundary accuracy of the models based on the worst-case scenario. While the standard Hausdorff distance focuses on the largest error, the HD95 metric used in this study reduces the effect of outlier distances by considering the 95th percentile of these distances (Taha & Hanbury, 2015).
By evaluating all these metrics together, both the regional overlap performance and boundary accuracy of the models were comprehensively analyzed.

4. Results

In this study, the performance of five deep learning models was evaluated on the BRISC 2025 dataset. The results of the quantitative evaluation metrics used to assess segmentation performance are summarized in Table 2.

Table 2. Quantitative segmentation results of the evaluated models on the BRISC 2025 test set.

MODEL	DICE	IOU	PRECISION	RECALL	HD95
FCN (Baseline)	0.7970	0.7142	0.8339	0.8375	8.5717
Swin U-Net	0.7257	0.6303	0.7806	0.7806	16.1159
U-Net	0.8466	0.7718	0.8951	0.8951	5.8975
DeepLabV3+	0.8575	0.7801	0.8457	0.9049	7.8367
Attention U-Net	0.8225	0.7497	0.8546	0.8604	9.6759

When the performance metrics presented in the table are examined, DeepLabV3+ achieves the highest performance with a Dice coefficient of 0.8575 and an HD95 value of 7.8367. U-Net, with a Dice value of 0.8466, demonstrates performance very close to DeepLabV3+. In terms of HD95 value, which measures the boundary accuracy of the models, DeepLabV3+ (7.8367) and U-Net (5.8975) exhibit lower error values compared to other models. No universally accepted threshold has been established for the HD95 metric, as it varies depending on the imaging modality and target structure size (Taha & Hanbury, 2015). In this study, HD95 values were calculated in pixels, where lower values indicate better boundary accuracy. The low HD95 value of the DeepLabV3+ model indicates that the model achieves superior performance in both regional overlap and boundary accuracy. The Transformer-based Swin U-Net model achieved the lowest performance with a Dice score of 0.7257 and an HD95 value of 16.1159. The Attention U-Net model performed below U-Net with a Dice value of 0.8225, while the FCN model outperformed the Swin U-Net model with a Dice value of 0.7970. The numerical results presented in the table revealed the overall performance of the models. Visual comparisons were also incorporated to better understand how accurately and consistently the models delineated tumor boundaries.
Accordingly, segmentation results of different models are presented on three representative images in the following figure:

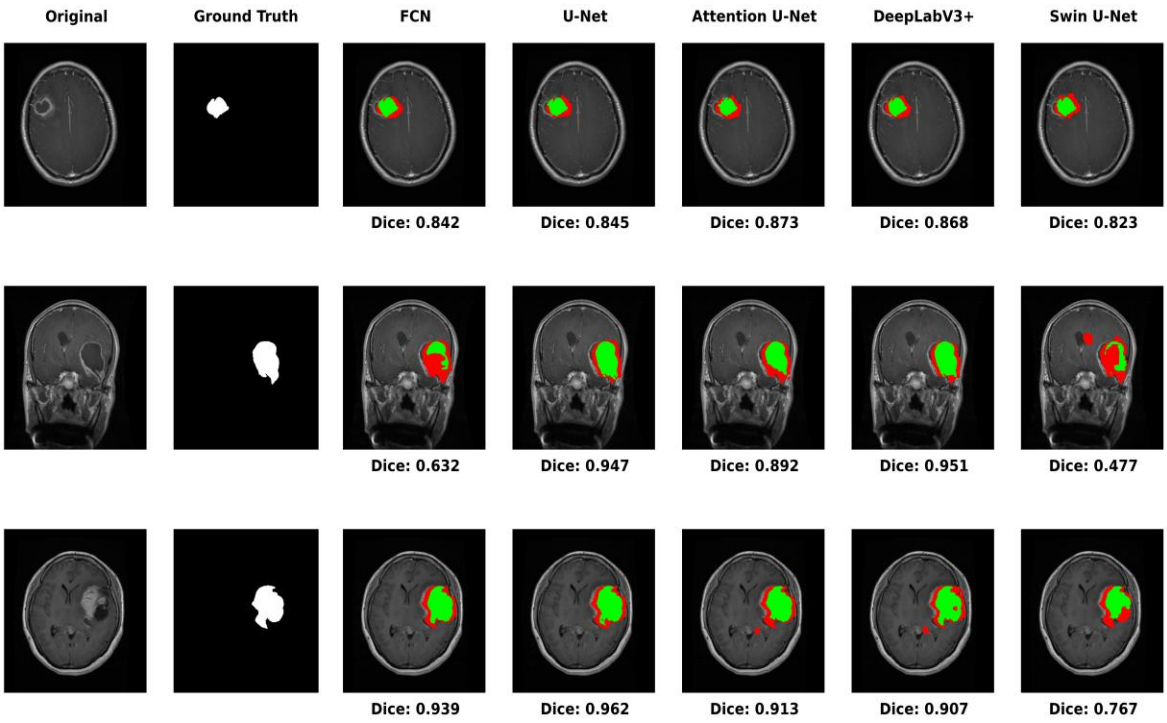

Figure 1. Color-coded segmentation comparison for three representative MRI images. Yellow: correctly segmented regions (A∩B); Red: missed tumor regions (A−B); Green: false positive predictions (B−A). A = Ground Truth, B = Model Prediction

As shown in Figure 1, a visual comparison of the segmentation outputs of the models indicates that DeepLabV3+ and U-Net models delineate tumor boundaries more accurately and consistently. These models produce smoother and more continuous segmentations, especially in regions with high tumor density.

Although the Attention U-Net model generally performs well, it occasionally fails to fully segment parts of the tumor region in some examples. The FCN model, despite its simpler architecture, identified the general tumor location but was less effective in capturing boundary details compared to other models.

The Transformer-based Swin U-Net model, while it can identify the general tumor location in some cases, struggles to accurately delineate boundary details and produces fragmented segmentations. The visual findings are consistent with the numerical results and further confirm that DeepLabV3+ and U-Net are superior in terms of segmentation performance.

5. Discussion

The findings of this study indicate that DeepLabV3+ and U-Net achieve higher performance in brain tumor segmentation compared to other models. The success of DeepLabV3+ suggests that its ASPP structure effectively processes multi-scale contextual information. This supports the critical role of multi-scale contextual feature extraction in segmentation performance.

The U-Net model achieved performance very close to DeepLabV3+ by preserving low-level spatial information through skip connections. A key finding is that the Transformer-based Swin U-Net performed below expectations, achieving the lowest performance with a Dice score of 0.7257 and a HD95 value of 16.1159.

The results highlight the high sensitivity of Transformer architectures to dataset characteristics. Compared to the stability exhibited by CNN-based models in terms of local feature extraction (inductive bias), the performance of models such as Swin U-Net is strongly correlated with the spatial distribution and variability of the dataset.

Previous studies have demonstrated that model performance is significantly influenced by dataset size and diversity. CNN-based approaches have been reported to provide effective and stable results even with limited datasets (Mostafa et al., 2023), whereas Transformer-based models tend to perform better on larger and more diverse datasets (Liu et al., 2021).

Table 3. Comparison of segmentation performance with related studies on the BRISC 2025 dataset

STUDY	ARCHITECTURE	METRIC	DATASET	INPUT SIZE	EVALUATION TYPE
Fateh et al. (2026)	SaberNet (Transformer)	Weighted mIoU: 80.6%	BRISC 2025	Not Reported	Overlap-based
Bhamboo et al. (2025)	Ensemble (UNet2.5D + AttUNet2.5D + TransUNet)	Dice: 0.798	BRISC 2025	128 × 128	Overlap-based
This Study	**DeepLabV3+ (CNN)**	**Dice: 0.8575, HD95: 7.8367**	**BRISC 2025**	**384 × 384**	**Overlap-based + Boundary-based**

Note: Results reported in the literature are based on different evaluation metrics (e.g., mIoU and Dice). Therefore, direct numerical comparison should be interpreted with caution.

These findings emphasize that both architectural design and dataset characteristics play a crucial role in model selection. These results were obtained under the specific dataset and experimental conditions used in this study. Training all models with identical hyperparameters ensured a fair and controlled comparison. Future studies exploring different hyperparameter configurations and model-specific optimization strategies may provide a more comprehensive understanding of their impact on performance. Furthermore, investigating ensemble or hybrid approaches that combine outputs from multiple architectures represents a promising research direction for improving segmentation performance.
This study has certain limitations. First, while all models shared identical hyperparameters (such as batch size and learning rate), CNN-based models (U-Net and DeepLabV3+) utilized ImageNet pre-trained backbones, whereas Attention U-Net and Swin U-Net were trained from scratch. This was deliberately done to evaluate their pure architectural learning capacities without transfer learning bias. The superior performance of DeepLabV3+ may be partially attributed to the robust feature extraction capabilities inherited from this pre-training. Therefore, the lower performance of the Transformer-based Swin U-Net emphasizes its inherent need for massive datasets or pre-trained weights to overcome the lack of inductive bias. Second, the use of identical hyperparameters for all models may not reflect their individual optimal configurations. Third, the evaluation was conducted on a single dataset, which may limit the generalizability of the findings.

6. Conclusion

The results demonstrate that DeepLabV3+ achieved the highest performance with a Dice score of 0.8575 and an HD95 value of 7.8367. This performance can be attributed to the high-resolution (384 x 384) input data and the hybrid function (Focal Dice + Focal Tversky). Furthermore, this suggests that the ASPP structure of the model, which can process multi-scale contextual information, is effective in this superior performance. The closest performance to DeepLabV3+ was achieved by the U-Net model, thanks to its skip connection mechanism that preserves spatial information. On the other hand, despite the attention mechanism used in the Attention U-Net model, it performed below the classical U-Net model. This may be attributed to the fact that the attention mechanism could not provide the expected level of learning under the current data structure. The performance of the Swin U-Net model indicates that it exhibits a structure more sensitive to the dataset characteristics. The results obtained in the study show that the model's performance is directly related to its architectural compatibility with the dataset. Therefore, these results suggest that Transformer-based structures respond more sensitively to changes in data distribution compared to CNN-based models.
This study provides a comprehensive comparison of CNN and Transformer-based models (FCN, U-Net, Attention U-Net, DeepLabV3+, Swin U-Net) on the BRISC 2025 dataset using the same dataset and fixed training conditions. The hybrid loss function (Focal Dice + Focal Tversky) and high-resolution input data (384 x 384) used enabled a more comprehensive evaluation of model performance. The findings highlight the impact of multi-scale contextual feature capture mechanisms on segmentation performance, and further indicate that Transformer-based models can be sensitive to dataset characteristics.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
ORCID iD: Yakup KUZHAN: https://orcid.org/0000-0003-1940-8279

References

[1] Abidin, Z. U., Naqvi, R. A., Haider, A., Kim, H. S., Jeong, D., & Lee, S. W. (2024). Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: A prospective survey. Frontiers in Bioengineering and Biotechnology, 12, 1392807. https://doi.org/10.3389/fbioe.2024.1392807
[2] Abraham, N., & Khan, N. M. (2019). A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI), 683-687.
[3] Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. Journal of Digital Imaging, 31(4), 449–459.
[4] Bhamboo, A. K., Mahala, S., Shekhawat, N. S., Kumar, Y., & Sharma, K. (2025). Trust-Refined U-Net Ensemble for Uncertainty-Aware Brain Tumor Segmentation. TechRxiv.
[5] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. arXiv preprint arXiv:2105.05537.
[6] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Proceedings of the European Conference on Computer Vision (ECCV), 801–818.
[7] Fateh, A., Rezvani, Y., Moayedi, S., et al. (2026). BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification. Scientific Data, 13(1), 361.
[8] Iqbal, A., Zafar, A., & Kim, J. (2023). Brain Tumor Segmentation from MRI Images Using Convolutional Neural Networks. Diagnostics, 13(16), 2650.
[9] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.
[10] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980-2988.
[11] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10012-10022.
[12] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431–3440.
[13] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV), 565–571.
[14] Mostafa, M. G., Hossain, M. S., & Islam, M. R. (2023). Brain Tumor MRI Segmentation Using Deep Learning: A Review. Diagnostics, 13(9), 1562.
[15] Ostrom, Q. T., et al. (2019). Brain Tumor Epidemiology: Consensus from the Brain Tumor Epidemiology Consortium. Neuro-Oncology, 21(4), 458–468.
[16] Pichaivel, M., Anbumani, G., Theivendren, P., & Gopal, M. (2022). An Overview of Brain Tumor. In: Brain Tumor, 1–14.
[17] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234–241.
[18] Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. International Workshop on Machine Learning in Medical Imaging, 379-387.
[19] Schlemper, J., Oktay, O., Schaap, M., et al. (2019). Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. Medical Image Analysis, 53, 197-207.
[20] Seshimo, H., Teramoto, A., Sugiura, M., & Fujita, H. (2024). Segmentation of Low-Grade Brain Tumors Using Mutual-Attention Model. Sensors, 24(23), 7576.
[21] Taha, A. A., & Hanbury, A. (2015). Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool. BMC Medical Imaging, 15(1), 29.
[22] Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.