Deep Architectures for Medical Image Segmentation: A Study on BraTS 2020

1st Aditya Jabade
Department of Applied Mathematics
aj3324@columbia.edu

2nd Sree Kuntamukkala
Department of Biomedical Engieering
sk5505@columbia.edu

***Abstract*—Brain tumor segmentation is an important task in neuro-oncology for diagnosis and treatment planning. Manual annotation is subjective and error-prone due to tumor heterogeneity. Accurate segmentation of tumor boundaries facilitates longitudinal tracking of tumor progression or recurrence. In this study, we explore the use of deep learning architectures to automate brain tumor segmentation using the BraTS 2020 dataset, a benchmark dataset composed of multimodal MRI scans (T1, T1c, T2, FLAIR) with expert-annotated tumor sub-regions. We implemented and trained four U-Net variants, including the baseline Vanilla U-Net, Attention enhanced Residual U-Net, UNet++, and the Swin U-Net, to evaluate their performance in identifying three clinically relevant sub-regions: the enhancing tumor, the tumor core, and edema. Each model was assessed using standard segmentation metrics such as Dice Similarity Coefficient and Jaccard Index. Our results provide insights into architecture effectiveness, establishing a foundation for further research in clinical-grade brain tumor segmentation.**

***Index Terms*—Brain tumor segmentation, U-Net, Swin Transformer, Attention U-Net, Deep Learning, BraTS 2020**

I. Background

Brain tumors, specifically gliomas, represent one of the most aggressive and deadly forms of cancer. According to the American Brain Tumor Association, over 700,000 people in the United States are currently living with a primary brain tumor, and approximately 88,000 new cases are expected to be diagnosed in 2025 alone, including 26,000 malignant tumors. Among these, glioblastoma multiforme (GBM) is the most common and lethal, with a median survival rate of only 15 months and a 5-year survival rate below 6.9% despite aggressive multi-modal therapy [1]. Diagnosis of brain tumors typically involves a combination of neuroimaging, clinical evaluation, and histopathological analysis. The gold standard is a surgical biopsy followed by genetic or molecular testing, which is invasive, carries procedural risks, and can delay treatment [7]. However, Magnetic Resonance Imaging (MRI) offers a non-invasive, less costlier and high-resolution modality for visualizing the tumor and its sub-regions, including enhancing tumor core, necrosis, and edema. Accurate interpretation of MRI scans is difficult because of the heterogeneity of gliomas, variability in MRI acquisition, and difference of opinion among radiologists. Manual segmentation of tumor regions is time-consuming and prone to high variability. To overcome these limitations, automated segmentation using deep learning is a promising approach.

In recent years, U-Net architectures—a class of Convolutional Neural Networks (CNNs)—have garnered significant attention in the medical imaging community due to their high accuracy in biomedical segmentation tasks. It has a symmetric encoder–decoder design with three primary components: the encoder, the bottleneck, and the decoder. The encoder progressively down-samples the input image while capturing semantic features through convolutional and pooling operations. The decoder performs up-sampling via transposed convolutions to reconstruct spatial resolution, enabling precise pixel-level segmentation. A key innovation of U-Net is the use of skip connections that directly link encoder and decoder layers at corresponding resolutions, allowing the model to combine coarse semantic information with fine-grained localization details [2].

Several enhancements to the original U-Net have been proposed to improve performance in complex medical imaging scenarios. Attention U-Net introduces attention gates into the skip connections, enabling the model to focus on the most relevant regions in the input image while suppressing irrelevant background features [3]. Residual U-Net incorporates residual connections within the encoder and decoder blocks, facilitating better gradient flow and enabling the training of deeper networks for improved feature representation [4].

UNet++ (also known as Nested U-Net) employs dense skip pathways and deep supervision to bridge the semantic gap between encoder and decoder features more effectively, enhancing feature reuse and segmentation accuracy [5]. More recently, Swin U-Net incorporates Swin Transformer blocks into the encoder, replacing traditional convolutions with self-attention mechanisms that capture long-range dependencies and contextual information throughout the image. These architectural innovations have expanded the applicability of U-Net variants to increasingly challenging medical segmentation tasks like brain tumor detection [6].

To evaluate these architectures, we use the BraTS 2020 dataset, a widely recognized benchmark for brain tumor segmentation. The BraTS data set provides preoperative multimodal MRI scans – including T1, T1 post-contrast (T1c), T2, and FLAIR modalities –from patients diagnosed with gliomas. Each case is annotated by experts with three clinically important tumor subregions: enhancing tumor (ET), edema (ED), and necrotic/non-enhancing tumor core (NCR/NET) [8]. The dataset's consistent labeling and thorough testing make

it the top choice for building and evaluating brain tumor segmentation models [9].

II. Problem Statement

This study aims to evaluate and compare the performance of various U-Net-based architectures for brain tumor segmentation using the BraTS dataset. Specifically, we investigate four models: the baseline Vanilla U-Net, an Attention-enhanced Residual U-Net that integrates attention mechanisms and skip connections, U-Net++ (or Dense U-Net) which employs nested and dense skip pathways, and Swin U-Net, which incorporates transformer-based modules for capturing global context. Our goal is to understand the strengths and limitations of each architecture in segmenting complex brain tumor structures.

III. Dataset and Pre-processing

As mentioned in the previous section this study uses the BraTS2020 dataset, a publicly available benchmark for brain tumor segmentation. It provides multi-modal MRI scans and expert-annotated segmentation masks for glioma patients, enabling robust evaluation of model performance across heterogeneous tumor structures [9].

- **Multimodal MRI Imaging Data**:
  - **T1**: Native T1-weighted (2D, 1–6 mm slice thickness)
  - **T1c**: Post-contrast T1-weighted (3D, 1 mm isotropic)
  - **T2**: T2-weighted (2D, 2–6 mm slice thickness)
  - **FLAIR**: Fluid-Attenuated Inversion Recovery (2D, 2–6 mm)

  Fig. 1 depicts one such sample across the above-mentioned modalities.
- **Segmentation Labels (provided by experts)**:
  - **Label 0**: Background
  - **Label 1**: Necrotic & non-enhancing tumor core (NCR/NET)
  - **Label 2**: Peritumoral edema (ED)
  - **Label 3**: Unused / No pixels
  - **Label 4**: Gadolinium-enhancing tumor (ET)

To ensure consistency across scans and improve training stability, the following pre-processing steps were applied:

- **Resizing**: All volumetric scans were resized to a uniform spatial resolution to facilitate batch processing.
- **Intensity normalization**: Pixel intensities were scaled to the range $[-1, 1]$ using `transforms.Normalize(mean=[0.5], std=[0.5])`, which maps original values from $[0, 1]$ to a centered distribution.

The dataset was split into training, validation, and testing subsets in a **70:20:10** ratio, respectively, from a total of **10,000** samples. A batch size of **16** was used during training.

IV. Architectures

A. Vanilla U-Net

The Vanilla U-Net is a simple yet popular encoder-decoder architecture originally designed for biomedical image segmentation [3]. It consists of a symmetric structure with a contracting path that captures context through convolution and max-pooling layers, and an expansive path for precise localization via upsampling and skip connections. These skip connections concatenate feature maps from the encoder to the decoder at corresponding levels, allowing the model to recover spatial information lost during downsampling.

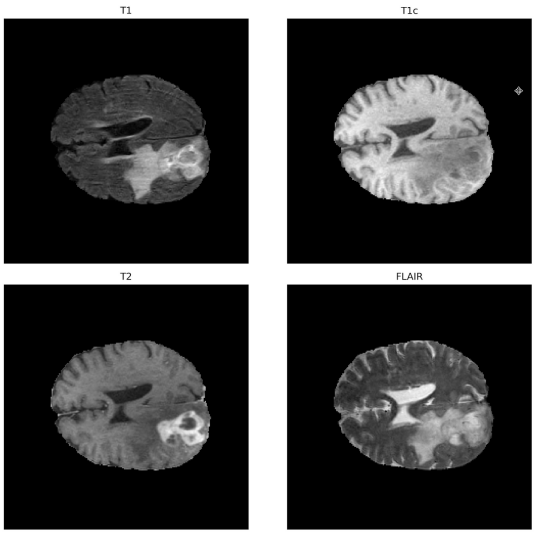

Fig. 1. Sample from BraTS2020 Dataset

B. U-Net++

U-Net++ is an advanced variant of U-Net that redesigns the skip connections using a series of nested and dense convolutional blocks [5]. Instead of directly connecting encoder and decoder features, U-Net++ gradually refines the feature maps through intermediate convolution layers before merging. This improves semantic consistency between encoder and decoder representations leading to more accurate and stable segmentation results. U-Net++ has demonstrated improved performance in medical image segmentation tasks, particularly where boundaries and fine structures are difficult to capture.

C. Attention enhanced Residual U-Net

The Attention UNet model is an enhanced version of the traditional U-Net, integrating attention mechanisms to improve feature fusion during the decoding process [3]. It follows the standard encoder-decoder architecture of U-Net, with a contracting path composed of convolutional and max-pooling layers (Down blocks) and an expansive path that performs upsampling and skip connection fusion (UpAttention blocks). The crux is in the AttentionBlock, which adaptively weights encoder features before concatenation based on their relevance to the decoder features, helping the model focus on more informative regions. Additionally, residual connections in the Double Convolutional blocks stabilize training and promote gradient flow. The use of bilinear upsampling followed by a 1×1 convolution in the upsampling path ensures smooth scaling, while attention gates suppress irrelevant features, enhancing segmentation accuracy.

D. Swin UNet

The Swin UNet model combines the plus points of the Swin Transformer backbone with a U-Net-inspired decoder to perform semantic segmentation on 2D medical images. The encoder uses a hierarchical vision transformer pretrained to extract multi-scale features, which are then progressively upsampled and combined with corresponding lower-level features through a series of custom Decoder Block modules. These decoder blocks perform learned upsampling via transposed convolutions and incorporate skip connections to preserve spatial information. The architecture ends with a final upsampling layer and a 1×1 convolution to produce pixel-wise class predictions. This structure leverages the use of the Swin Transformer backbone for segmentation tasks has been shown to enhance performance on medical datasets by capturing both local and global context efficiently. We chose not to include architectural figures for the models discussed above due to two key reasons: their large size made them difficult to fit within the page layout, and we were constrained by the 4-page limit for the report. However we have those images in our presentation slides.

V. Training & Validation

For training of the proposed models, we used the Adam optimizer with a learning rate of 0.001 and trained the network for 30 epochs. The loss function used was Cross Entropy Loss, which is well-suited for multi-class segmentation tasks as it effectively penalizes incorrect class predictions. To evaluate segmentation performance, we calculated both the multi-class Dice Score and multi-class Jaccard Index, which measure the overlap between predicted and ground truth segmentation masks. These metrics were chosen because they are well suited for our multi-class segmentation problem. Throughout training and validation, we plotted the loss curves, Dice curves, and Jaccard curves to check for overfitting or underfitting trends. Additionally, a confusion matrix was generated for each model to visualize class-wise prediction performance and detect common misclassifications.

A. Results

1) Quantitative: The loss curves in Figure 1 show that Vanilla UNet converges slowly with high final losses. U-Net++ and Attention UNet achieve faster convergence and lower losses, with Attention UNet showing more stable validation performance. Swin UNet reaches the lowest losses overall, despite minor fluctuations (possibly due to imbalance in the BRATS2020 Dataset). The Dice score curves (Figure 2) show that advanced architectures outperform the baseline Vanilla UNet in segmentation performance. For the same number of epochs, all the three U-Net++, Attention enhanced Residual UNet and the Swin UNet reach to Dice scores of 0.9 where as the Vanilla UNet is just about 0.84. A similar observation is noted for the Jaccard score. The reason behind superior performance of the advanced UNets are most probably due to: enhanced skip connections and attention mechanisms for the UNet++ and the Attention UNet respectively and the transformer-based attention which captures complex features in the Swin UNett. Table I presents a comparative analysis of four segmentation models—UNet, UNet++, Attention UNet, and Swin UNet—evaluated on multiple performance metrics. Among all, the Attention UNet achieved the highest validation Dice score (**0.893**) and Jaccard score (**0.812**), indicating superior segmentation quality. This improvement can be attributed to the use of attention gates, which help the model focus on relevant spatial features while suppressing noise. Swin UNet also performed competitively, benefiting from transformer-based feature extraction that captures long-range dependencies; however, its slight drop in Dice score suggests that it may require more data or tuning to outperform convolution-based models in this task. UNet++ outperformed the Vanilla UNet due to its nested and dense skip connections, which allow richer feature reuse and better gradient flow. While the Vanilla UNet achieved high accuracy, its higher loss values suggest poorer class-wise precision, possibly due to limited contextual awareness (absence of attention). Overall, the results demonstrate the advantages of architectural enhancements like attention mechanisms and transformer backbones in improving segmentation performance on complex medical images.

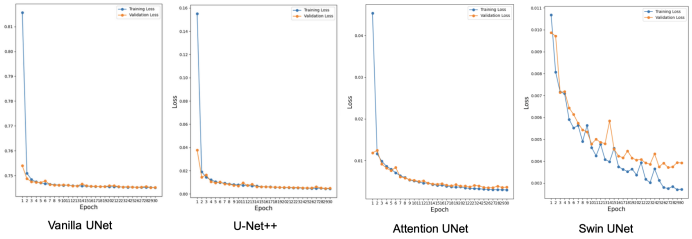

Fig. 2. Loss curves for Vanilla UNet, UNet++, Attention UNet, and Swin UNet

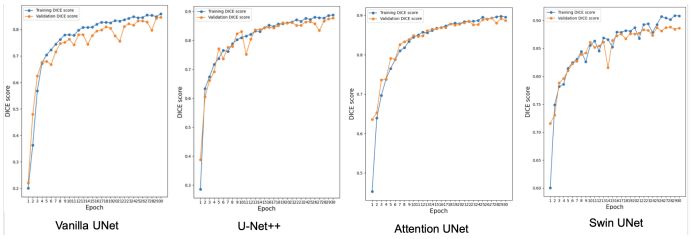

Fig. 3. Dice Score curves for Vanilla UNet, UNet++, Attention UNet and Swin UNet

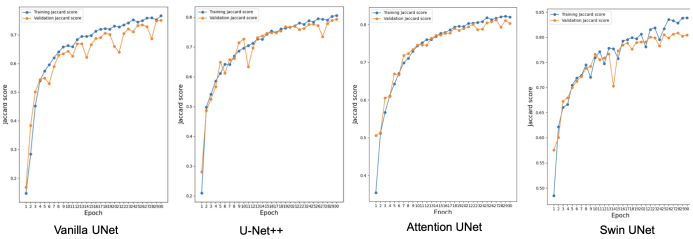

Fig. 4. Jaccard Score curves for Vanilla UNet, UNet++, Attention UNet and Swin UNet

TABLE I
COMPARISON OF SEGMENTATION MODEL PERFORMANCE

Model	Best Epoch	Val Dice	Val Jaccard	Test Acc	Train Loss	Val Loss	Test Loss
UNet	30	0.846	0.750	0.9984	0.745	0.745	0.74507
UNet++	30	0.8872	0.7928	0.9985	0.0045	0.0049	0.0037
Attention UNet	27	0.893	0.812	0.99854	0.0030	0.0035	0.0036
Swin UNet	28	0.888	0.808	0.9985	0.0020	0.0037	0.0037

2) Qualitative: Fig. 5 depicts the output of the Vanilla UNet on a sample testing slice. Fig. 6 presents qualitative segmentation results overlaid on FLAIR MRI slices, comparing the ground truth (green) and predicted tumor regions (red) for four models: Vanilla UNet, U-Net++, Attention UNet, and Swin UNet. The Vanilla UNet prediction shows a slight mismatch at the tumor boundaries (visible only on zooming). U-Net++, Attention UNet and the Swin UNet exhibit better alignment with the ground truth, capturing fine tumor borders. These visual results align with the quantitative metrics, further validating the performance differences among the models.

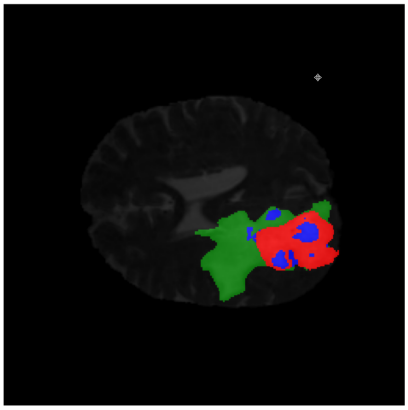

Fig. 5. Sample Output of Vanilla UNet

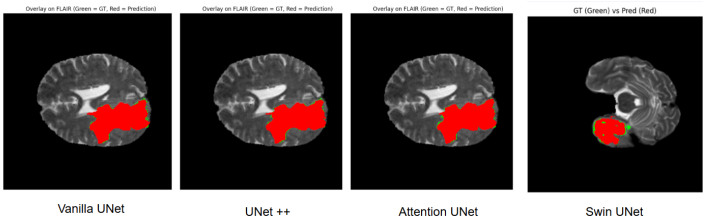

Fig. 6. Qualitative Analysis Vanilla UNet, UNet++, Attention UNet and Swin UNet

VI. INFERENCES & DISCUSSION

The comparative analysis of Vanilla UNet, U-Net++, Attention UNet, and Swin UNet revealed key insights into their segmentation capabilities on the BraTS 2020 dataset. From the loss, Dice, and Jaccard curves, it is evident that more advanced architectures exhibit better convergence behavior and generalization. Vanilla UNet showed limited learning capacity, with higher losses and relatively lower Dice and Jaccard scores. In contrast, U-Net++ and Attention UNet demonstrated significant performance improvements by leveraging dense skip connections and attention mechanisms, respectively. These models showed faster convergence, lower losses, and higher segmentation accuracy, with more stable validation metrics across epochs. Attention enhanced residual UNet and the Swin UNet jointly performed the best both quantitatively and qualitatively.

Overall, the experiments highlight the importance of architectural advancements in medical image segmentation tasks. Models with enhanced feature fusion (U-Net++), attention mechanisms (Attention UNet), and transformer-based encoding (Swin UNet) provide significant advantages over traditional CNN-based architectures. These findings suggest that future research should continue to explore hybrid and transformer-based models, particularly for complex volumetric medical datasets.

However, it is worth noting that this study was conducted on the well-structured and widely used BraTS 2020 dataset, which may not fully reflect the variability and noise present in real-world clinical data. Applying these models to real-world datasets would provide a more rigorous test of their generalizability. Moreover, since the top performing models achieved comparable performance in terms of loss and accuracy on BraTS, evaluating them on more diverse, real-world data could help draw sharper distinctions and yield more clinically meaningful insights.

VII. FUTURE DIRECTIONS

Moving forward, the project can be expanded in several directions. One of the things that we plan to do is to explore MedSegMamba [10]. This a hybrid model that combines 3D CNNs with Mamba blocks using Selective Scan 3D (SS3D) to efficiently model long-range dependencies in volumetric MRI segmentation. Additionally, we plan on benchmarking other UNet variants such as nnUNet, R2 UNet, and Inception UNet to provide a more comprehensive comparative analysis. These extensions aim to deepen understanding of architectural trade-offs and improve performance in real-world medical imaging tasks.

ACKNOWLEDGMENT

We would like to sincerely thank Professor Jia Guo for his invaluable guidance and support throughout the BMEN 4460 course. We also extend our gratitude to the teaching assistants for their continuous help, timely feedback, and dedication, which played a crucial role in the successful completion of this project.

REFERENCES

[1] Ostrom, Quinn T., et al. "CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019." Neuro-oncology24.Supplement_5 (2022): v1-v95
[2] https://arxiv.org/abs/1505.04597
[3] https://arxiv.org/abs/1804.03999
[4] https://arxiv.org/abs/1711.10684
[5] https://arxiv.org/abs/1807.10165
[6] https://arxiv.org/abs/2105.05537
[7] Menze, Bjoern H., et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." IEEE transactions on medical imaging 34.10 (2014): 1993-2024.

[8] Bakas, Spyridon, et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge." arXiv preprint arXiv:1811.02629 (2018).
[9] https://www.kaggle.com/datasets/andrewmvd/brain-tumor-segmentation-in-mri-brats-2015.
[10] https://arxiv.org/abs/2409.08307