Brain Tumor Segmentation

A Deep Learning Approach

Federico Cesare Cattó
Master's Degree in Data Science
Università degli Studi di Milano-Bicocca

August 27, 2025

Contents

1 Introduction

Brain tumors, particularly gliomas, are aggressive neoplasms that affect the central nervous system and require accurate and timely diagnosis. Magnetic Resonance Imaging (MRI) is the primary non-invasive imaging technique used for tumor detection, treatment planning, and monitoring. However, manual segmentation of brain tumors in MRI scans is labor-intensive, time-consuming, and highly dependent on the expertise of radiologists. This process is also prone to inter-observer variability, making it inconsistent across different clinicians.

Automating the segmentation of brain tumor regions through Artificial Intelligence (AI) and Deep Learning (DL) has the potential to improve clinical workflows by providing objective, reproducible, and efficient delineation of tumor boundaries. Among various deep learning architectures, convolutional neural networks (CNNs) and in particular the U-Net architecture have demonstrated high effectiveness in medical image segmentation tasks.

This project aims to develop a deep learning model capable of segmenting brain tumor subregions—namely the Enhancing Tumor (ET), Whole Tumor (WT), and Tumor Core (TC)—using multi-modal MRI data from the BraTS dataset. Our goal is to optimize segmentation accuracy through careful preprocessing, model design, and evaluation, ultimately contributing to more precise and faster clinical decision-making.

2 Dataset

The dataset used in this project is the **BraTS (Brain Tumor Segmentation)** dataset, a globally recognized benchmark in the field of medical image segmentation. BraTS provides multi-institutional, multi-modal MRI scans that are anonymized and annotated by clinical experts, ensuring high-quality ground truth for training and evaluation.

2.1 Overview

The BraTS dataset used in this project consists of a total of 484 labeled training cases and 300 unlabeled test cases. Each patient sample includes four co-registered MRI volumes, each with a spatial resolution of $240 \times 240 \times 155$ voxels, resulting in high-resolution 3D data.

The dataset provides four MRI modalities per patient, which capture different tissue characteristics:

- **FLAIR (Fluid-Attenuated Inversion Recovery)**: highlights edema and fluid accumulation, especially in areas surrounding the tumor.
- **T1-weighted (T1w)**: offers detailed anatomical structure of the brain.
- **T1 with gadolinium contrast (T1gd)**: enhances the visibility of active tumor regions by highlighting areas where the contrast agent accumulates.
- **T2-weighted (T2w)**: emphasizes fluid-tissue contrast, aiding in the distinction between normal and abnormal tissue.

These four modalities, when combined, give the model a comprehensive view of the tumor's morphology and environment, significantly improving the learning of complex spatial patterns.

2.2 Segmentation Labels

The ground truth segmentation masks provided by the BraTS dataset use a set of class labels to identify different tumor components. Specifically, Label 1 corresponds to the non-enhancing tumor core (NET), Label 2 denotes the peritumoral edema, and Label 4 identifies the enhancing tumor (ET). These labels are not only important individually, but are also commonly grouped into broader tumor subregions that are clinically meaningful. The Enhancing Tumor (ET) corresponds to Label 4, the Tumor Core (TC) includes both Labels 1 and 4, and the Whole Tumor (WT) encompasses all three labels: 1, 2, and 4.

Working with this dataset presents several challenges. First, the MRI volumes are high-resolution and contain four separate imaging modalities per case, leading to substantial memory and processing demands. Second, there is a significant class imbalance: the vast majority of pixels represent healthy tissue or background, with tumor regions comprising only a small fraction of the data. This imbalance can bias the model toward predicting the majority class unless properly addressed. Finally, the tumors themselves are highly heterogeneous—they can vary greatly in size, shape, location, and intensity across patients. This makes the segmentation task particularly complex, requiring the model to generalize well to a wide range of visual patterns.

3 Preprocessing

Preprocessing plays a fundamental role in ensuring that the input data is standardized, balanced, and ready for deep learning training. Given the complexity and size of the BraTS dataset, specific preprocessing steps were applied to address both computational and modeling challenges.

3.1 Slice Extraction

Since the MRI volumes are 3D, we extracted only the 2D slices that contain visible tumor tissue. This approach reduces unnecessary data and ensures the model focuses on informative regions. A tracking system was implemented to preserve patient identity during the train-validation split, preventing data leakage.

3.2 Mask Recoding

The original segmentation masks use labels {0, 1, 2, 4}. For model compatibility and training efficiency, they were recoded into a continuous range {0, 1, 2, 3}, where:

- 0: Background
- 1: Non-enhancing Tumor Core
- 2: Peritumoral Edema
- 3: Enhancing Tumor

3.3 Normalization

To ensure consistent intensity ranges across patients and modalities, z-score normalization was applied *per modality*. Each modality was standardized individually by subtracting the mean and dividing by the standard deviation of non-zero voxels.

3.4 Cropping

To reduce computational load and eliminate irrelevant background regions, each 2D MRI slice was cropped around the brain region. This was done by identifying the smallest rectangular area that contains all non-zero pixels (i.e., actual brain tissue), effectively removing empty black borders that contain no useful information. By narrowing the focus to the region of interest, this operation preserves the anatomical content of the scan while reducing input size, which accelerates training and helps the model converge faster and more reliably.

3.5 Data Augmentation

To address the severe class imbalance present in the dataset—where background pixels vastly outnumber tumor pixels—and to improve the model's ability to generalize, we applied various data augmentation techniques. These techniques were designed both to increase the diversity of the training data and to simulate the variability typically encountered in real clinical settings.

We experimented with a wide range of augmentations, including spatial transformations (such as random rotations, flips, and elastic deformations), and non-spatial transformations (such as Gaussian noise, brightness and contrast adjustments, and gamma correction). However, during testing, we observed that spatial augmentations often introduced unrealistic anatomical distortions in the brain structure, which negatively affected model performance. For this reason, we decided to exclude geometric transformations from the final pipeline.

The final set of augmentations included only non-spatial modifications. These augmentations successfully mimicked acquisition variability without compromising anatomical integrity and helped the model become more robust to differences in intensity and image quality across patients and scanners.

4 Model Architecture

The segmentation task was addressed using a **2D U-Net architecture**, a deep learning model widely adopted in medical image segmentation due to its ability to combine global context with fine localization. The U-Net follows an encoder–decoder structure with characteristic skip connections that directly transfer spatial information from the encoder to the decoder, ensuring accurate boundary detection.

The model input consists of MRI slices of size 240×240 with four channels, corresponding to the different modalities (FLAIR, T1w, T1gd, and T2w). The output is a pixel-wise classification into four categories: background, non-enhancing tumor core, peritumoral edema, and enhancing tumor.

Encoder Path

The encoder path is composed of four convolutional blocks. Each block includes two convolutional layers with 3×3 kernels, batch normalization, and ReLU activations, followed by a 2×2 max pooling operation. The number of filters increases progressively across the blocks: 64, 128, 256, and 512. This design allows the network to gradually capture higher-level semantic features while reducing the spatial resolution.

Bottleneck

At the bottom of the network, the bottleneck layer consists of two convolutional layers with 1024 filters, providing a compressed but highly informative representation of the input features before reconstruction begins.

Decoder Path

The decoder path mirrors the encoder but with upsampling operations. Each block applies a transposed convolution (2×2, stride 2) to increase the spatial resolution, followed by concatenation with the corresponding encoder feature maps through skip connections. Two additional convolutional layers with batch normalization and ReLU activations are then applied. The number of filters decreases symmetrically across the decoder: 512, 256, 128, and 64.

Output Layer

Finally, a 1×1 convolution with softmax activation is used to assign a class probability to each pixel, resulting in segmentation masks with four output classes (background and three tumor subregions).

This architecture preserves the detailed spatial information of the tumor regions while simultaneously leveraging deep hierarchical features, making it highly suitable for brain tumor segmentation tasks.

Evaluation: Dice Metrics

To evaluate segmentation performance, we adopted the **Dice coefficient**, a widely used metric in medical image analysis. The Dice score is defined as the harmonic mean of precision and recall, measuring the degree of overlap between the predicted segmentation and the ground truth annotation. A higher Dice score indicates a better match between the two, with a perfect segmentation yielding a Dice score of 1.

In practice, the Dice score is particularly well suited to tasks with severe class imbalance, such as brain tumor segmentation, since it directly accounts for both false positives and false negatives. This makes it more reliable than standard accuracy, which can be misleading when background voxels dominate the dataset.

Training was optimized by minimizing the **Dice loss**, defined as $1 - \text{Dice Score}$. By doing so, the model learns to maximize the overlap between predicted and true tumor regions while penalizing incorrect classifications. This loss function proved effective in addressing the imbalance between tumor and non-tumor voxels, thereby improving the robustness of the segmentation results.

5 Training and Evaluation

During the experimentation phase, several configurations of preprocessing and augmentation were tested to assess their impact on segmentation performance. The main goal was to identify the pipeline that achieved the best balance between accuracy and generalization.

Initial attempts suffered from **data leakage**, where slices from the same patient appeared in both training and validation sets. This issue led to artificially inflated performance. Once corrected by carefully tracking patient IDs during dataset splitting, more realistic and reliable evaluation was achieved.

Subsequent experiments explored different preprocessing strategies. Normalization by modality and cropping around the brain region were introduced to improve consistency and reduce irrelevant background information. Further experiments added data augmentation techniques, such as rotations and noise injection, in order to address class imbalance and improve robustness.

The combined preprocessing pipeline—featuring modality-wise normalization, targeted non-spatial augmentation, and cropping—produced the best results, reaching an average **Dice Score of 0.53** with a corresponding **Dice Loss of 0.66**. These values reflect moderate overlap between predicted and ground truth tumor regions, highlighting both the strengths and limitations of the baseline U-Net model.

6 U-Net Experimentation & Results

The results highlight the segmentation performance of the baseline U-Net as well as the experimental variants, Attention U-Net and Residual U-Net.

The baseline U-Net achieved an average **Dice Score of 0.53**, which indicates moderate overlap between the predicted tumor regions and the expert-annotated ground truth. Visual inspection of the results confirmed that the model was able to detect the general location and shape of tumors, but often struggled with precise boundary delineation, especially for smaller or irregularly shaped regions.

The **Attention U-Net** variant improved performance, achieving a **Dice Score of 0.64**. By integrating attention gates into the skip connections, this model was better able to focus on the most relevant features while suppressing irrelevant activations, leading to more accurate segmentation masks.

The **Residual U-Net** obtained a **Dice Score of 0.58**, showing only a marginal improvement over the baseline. Although residual connections and dropout provided better regularization, this variant did not outperform the Attention U-Net in terms of accuracy.

Qualitative results further illustrate the differences among models. For most cases, the baseline U-Net detected the tumor regions but underestimated their extent, while the Attention U-Net produced more refined segmentations that closely aligned with the ground truth. The Residual U-Net performed adequately but tended to be less consistent across patients.

Overall, the **Attention U-Net** was selected as the best-performing architecture in this

project.

7 Discussion

The experimental results provided several insights into the challenges and limitations of brain tumor segmentation using deep learning. First, the preprocessing steps—such as modality-wise normalization, cropping, and non-spatial augmentation—did not lead to a substantial improvement in segmentation accuracy. While they were essential to prepare the data correctly and prevent data leakage, their impact on the Dice Score remained limited.

Second, the model comparison revealed that not all architectural modifications yield clear benefits. The Residual U-Net introduced residual connections and dropout, which improved regularization, but overall performance was not significantly better than the baseline U-Net. In contrast, the Attention U-Net consistently outperformed both models by effectively focusing on the most relevant features, resulting in more precise segmentation boundaries.

These findings confirm the importance of carefully evaluating each design choice rather than assuming that additional complexity automatically translates into better results. Moreover, the experiments highlighted the difficulty of achieving high Dice Scores in this task due to factors such as class imbalance, limited computational resources, and the intrinsic variability of tumor structures across patients.

Overall, the discussion emphasizes that while the Attention U-Net represents a step forward compared to the baseline, there is still room for improvement, particularly through more advanced architectures and enhanced preprocessing strategies.

8 Conclusion

In conclusion, this project demonstrated the application of deep learning techniques for brain tumor segmentation using the BraTS dataset. By implementing a U-Net architecture and exploring its variants, we were able to assess the strengths and weaknesses of different approaches. The baseline U-Net provided moderate results, while the Attention U-Net achieved the best overall performance, confirming the benefit of incorporating attention mechanisms in medical image segmentation tasks.

Although the obtained Dice Scores show that there is still significant room for improvement, the experiments validated the potential of AI-based approaches to support clinical decision-making. More advanced models, improved preprocessing pipelines, and larger computational resources could further enhance performance in future work.

9 Clinical Impact & Future Directions

The clinical relevance of this project lies in its ability to enhance the precision and efficiency of brain tumor analysis. Automated segmentation provides a consistent and objective tool that can support radiologists in diagnosis and treatment planning. By reducing the time required for manual annotation and minimizing inter-observer variability,

AI-based methods have the potential to streamline clinical workflows and improve patient outcomes.

Looking ahead, several directions could strengthen the robustness and applicability of this work. First, validation on external clinical datasets is essential to assess the generalizability of the models beyond the BraTS benchmark. Second, more advanced data augmentation and class balancing strategies could address the severe imbalance between tumor and non-tumor voxels. Third, moving from 2D slice-based models to full 3D architectures would allow the network to capture richer spatial context, likely improving segmentation accuracy. Finally, the integration of more sophisticated architectures, such as transformer-based models, could further enhance performance by better capturing global dependencies within MRI scans.

10 Project Limitations

Despite the promising results, this project faced several limitations that constrained both model performance and scope. The first major limitation was **computational resources**. The available RAM and GPU capacity were insufficient to train on large-scale data or to implement full 3D models, which limited the exploration of architectures capable of leveraging volumetric information. As a consequence, experiments were restricted to 2D slice-based approaches, which do not fully capture the spatial continuity of tumor structures.

Another limitation was the **reduced ability to perform extensive hyperparameter tuning**. Long training times prevented systematic optimization, which may have hindered the performance of all tested models. In addition, the relatively modest results reflect the **lack of advanced medical imaging expertise** at the project stage, which could have influenced decisions in preprocessing and evaluation strategies.

Finally, the models were trained and evaluated only on the BraTS dataset, meaning that their **generalizability to external clinical data** remains unverified. Without validation on real-world hospital data, it is difficult to fully assess their robustness in clinical practice.