BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Explainable AI for Enhanced Brain Tumor Segmentation in Multi-modal MRI Scans

Rafid Nahiyan Farabi
190041131
Tasnim Ferdous
190041139
H. M. Shadid Reza Chowdhury
190041219

Department of Computer Science and Engineering
Islamic University of Technology
June, 2024

Explainable AI for Enhanced Brain Tumor Segmentation in Multi-modal MRI Scans

Rafid Nahiyan Farabi
190041131
Tasnim Ferdous
190041139
H. M. Shadid Reza Chowdhury
190041219

Department of Computer Science and Engineering
Islamic University of Technology
June, 2024

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Rafid Nahiyan Farabi**, **Tasnim Ferdous**, and **H. M. Shadid Reza Chowdhury** under the supervision of **Dr. Md. Azam Hossain**, Associate Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Dr. Md. Azam Hossain
Associate Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: June 04, 2024

Rafid Nahiyan Farabi
Student ID: 190041131
Date: June 04, 2024

Tasnim Ferdous
Student ID: 190041139
Date: June 04, 2024

H. M. Shadid Reza Chowdhury
Student ID: 190041219
Date: June 04, 2024

Dedicated to our parents, whose unwavering support and guidance have been our strength in both good times and bad

Contents

List of Figures

List of Tables

Acknowledgement

We are deeply indebted to our supervisor, **Dr. Azam Hossain**, for his boundless patience, insightful critiques, and timely advice. His expertise and unwavering encouragement have been pivotal in the successful completion of this thesis. His invaluable support and motivational talks, often disguised as time management lectures, have been a constant source of inspiration. His steadfast belief in our abilities kept us moving forward, even when our faith wavered.

We extend our sincere gratitude to **Dr. Iqram Hussain** from Well Cornell Medicine, Cornell University, NY, USA, for his significant contributions to this research. His expert guidance and valuable insights have greatly enriched our work.

A special note of gratitude goes to our parents, whose love and support have been our anchor. Thank you for always being there, for listening to my lengthy explanations of my research, and for your unwavering patience during my journey. Your belief in me has been a source of great strength.

To all of you, I extend my deepest appreciation and gratitude. Your support has been invaluable, and I promise to navigate fewer detours through the realm of streaming services in the future.

Abstract

In the recent years, Explainable AI (XAI) and convolutionl neural networks (CNNs) have achieved significant breakthrough in the medical image analysis. This thesis investigates the application of deep learning models for the segmentation of brain tumors in multi-modal MRI scans, incorporating Explainable AI (XAI) methodologies to enhance interpretability. Using the BraTS Challenge 2020 dataset, this study implements and compares various deep learning architectures, including U-Net, UnetR, swin U-Net to segment brain tumors effectively. The research addresses critical challenges such as the complexity of MRI image preprocessing, limited availability of annotated data, and high computational requirements. To ensure robust and reliable model performance, advanced loss functions and data augmentation techniques including flipping, rotation, scaling, brightness adjustment, and elastic deformation are used. The core contribution of this work lies in the integration of XAI techniques like GradCAM, Feature Ablation, and Kernel SHAP to elucidate the decision-making processes of the neural networks. This approach aims to bridge the gap between model predictions and clinical interpretability, providing heatmaps and visual explanations that highlight the most influential regions in the MRI scans. The results demonstrate the efficacy of the proposed models in accurately segmenting tumor regions, with the XAI methods offering valuable insights into model behavior and reliability. This research contributes to the development of more transparent and trustworthy AI systems in brain tumor segmentation. The findings hold significant implications for enhancing clinical decision-making and personalized treatment planning for brain tumor patients.

Chapter 1

Introduction

1.1 Overview

In the field of medical diagnosis and treatment cancerous brain and Central Nervous System (CNS) tumors pose a significant hurdle [49]. This necessitating precise and accurate segmentation of tumors in MRI scans for effective clinical decision-making. Traditional methods of brain tumor segmentation are labor-intensive and subject to variability, prompting the need for automated solutions [46]. Recent advancements in deep learning have shown promise in medical image analysis, particularly in the segmentation of brain tumors. This thesis explores the application of deep learning models, namely U-Net [34], U-Net++ [48], and Swin U-Net [6], in the segmentation of brain tumors from multi-modal MRI images, coupled with Explainable AI (XAI) [15] techniques to enhance the interpretability and reliability of the segmentation results.

1.2 Motivations and Scope

Brain tumors account for 85% to 90% of all primary Central Nervous System (CNS) tumors [8]. In 2020, an estimated 251,329 people worldwide died from primary cancerous brain and CNS tumors. After diagnosis with brain tumor, only 21% of people over the age of 40 survive in the next 5 years [8]. There is a critical need for accurate and efficient segmentation of brain tumors to assist in diagnosis, treatment planning, and monitoring. Automated segmentation reduces the time and effort required by radiologists and provides consistency in the analysis [39]. Despite the progress made by existing deep learning models, challenges remain in achieving high accuracy across diverse datasets and ensuring the interpretability of the results. This study aims to

address these challenges by leveraging the BraTS 2020 [2] dataset, which provides a comprehensive collection of multi-modal MRI images, to train and evaluate the performance of advanced segmentation models. Furthermore, the incorporation of XAI techniques seeks to provide insights into the decision-making process of these models, thereby increasing their trustworthiness in clinical settings.

1.3 Problem Statement

The primary problem addressed in this thesis is the accurate and interpretable segmentation of brain tumors in multi-modal MRI images using deep learning models. Specifically, the research focuses on improving segmentation accuracy and enhancing the explainability of the models' predictions to facilitate their adoption in clinical practice.

1.4 Research Challenges

Brain tumor segmentation presents numerous challenges that complicate the accurate delineation of tumor regions in medical imaging. One significant challenge is location uncertainty. Brain tumors, such as gliomas, can occur anywhere in the brain due to the widespread distribution of glial cells [28]. This unpredictability in location makes it difficult to create models that can accurately identify tumor regions across different patients. Additionally, the morphological uncertainty is another obstacle, with brain tumors exhibiting significant variations in shape and size. The presence of surrounding edema, or fluid-filled tissue, further complicates the process, as it does not provide consistent indicators of the tumor's shape.

Low contrast in MRI images also poses a major challenge. These images, commonly used for brain tumor detection, often suffer from low contrast and image quality, making it difficult to distinguish tumor boundaries from healthy tissue, leading to less precise segmentation. **Annotation bias** further complicates matters, as the process of manually labeling tumor regions in medical images is subjective, relying heavily on the annotator's expertise. This can result in inconsistent annotations, which affect the training of segmentation algorithms. Additionally, the imbalanced issue in tumor segmentation refers to the disproportionate representation of different tumor regions within the data, skewing the performance of data-driven algorithms.

Further challenges include the variability in tumor appearance and location, requiring models that can generalize across different tumor types and MRI modalities. The

need for large and diverse annotated datasets is also critical, as well as the computational complexity involved in training deep learning models. Lastly, the difficulty in interpreting the models' decisions underscores the need for methods that can elucidate the inner workings of these models. Addressing these challenges requires the development of robust and generalized models capable of performing well across various brain tumor types and MRI modalities.

1.5 Contribution

TThis thesis advances the field of medical image analysis, specifically in brain tumor segmentation, through the following key contributions:

1. A comparative analysis of three advanced deep learning models—U-Net [34], U-Net++ [48], and U-Net R [21] applied to brain tumor segmentation using the BraTS 2020 dataset.
2. The novel application of multiple XAI techniques, including GradCAM [37], feature ablation, and Kernel SHAP [9], to provide visual and quantitative explanations of the segmentation results.
3. A evaluation framework that assesses the effectiveness of aforementioned models and XAI techniques, improving the accuracy and interpretability of brain tumor segmentation.
4. Critcal insights into the strengths and limitations of each model, contributing to the development of more reliable and interpretable medical image analysis tools.

1.6 Organization

The remainder of this thesis is organized as follows:

- Chapter 2 reviews related works in the field of brain tumor segmentation and explainable AI, highlighting key studies and existing methodologies.
- Chapter 3 details the materials and methods used, including an overview of the dataset, data preparation and augmentation techniques, loss functions, and evaluation metrics.
- Chapter 4 presents the experimental setup and results, comparing the performance of the different models and analyzing the effectiveness of the XAI tech-

niques.

- Chapter 5 discusses the results, their implications, and potential areas for future research.
- Chapter 6 concludes the thesis, summarizing the findings and contributions of the research.

By following this structure, the thesis aims to provide a comprehensive and coherent exploration of the application of deep learning and XAI in brain tumor segmentation, ultimately contributing to the advancement of medical image analysis.

Chapter 2

Related Works

Biomedical image segmentation is a crucial task in medical imaging, enabling the precise identification and localization of anatomical structures. Various neural network architectures have been developed to tackle this problem, among which the U-Net and its variants have shown remarkable success. This literature review provides a comprehensive overview of the original U-Net [34], 2D U-Net, and Swin U-Net [6], highlighting their architectures, performance, and implementation details.

2.1 U-Net

The U-Net [34] consists of a series of convolutional layers followed by max-pooling layers in the contracting path, and up-convolutional layers in the expanding path. The skip connections between the contracting and expanding paths allow for the retention of high-resolution features, which are crucial for precise segmentation. The U-Net consists of a series of convolutional layers followed by max-pooling layers in the contracting path, and up-convolutional layers in the expanding path. The skip connections between the contracting and expanding paths allow for the retention of high-resolution features, which are crucial for precise segmentation [34]. U-Net has demonstrated superior performance in biomedical segmentation tasks. It outperformed prior methods such as sliding-window convolutional networks on the ISBI challenge for segmentation of neuronal structures in electron microscopy images. U-Net also excelled in the ISBI cell tracking challenge, achieving a significantly higher Intersection over Union (IOU) score compared to other algorithms [34].

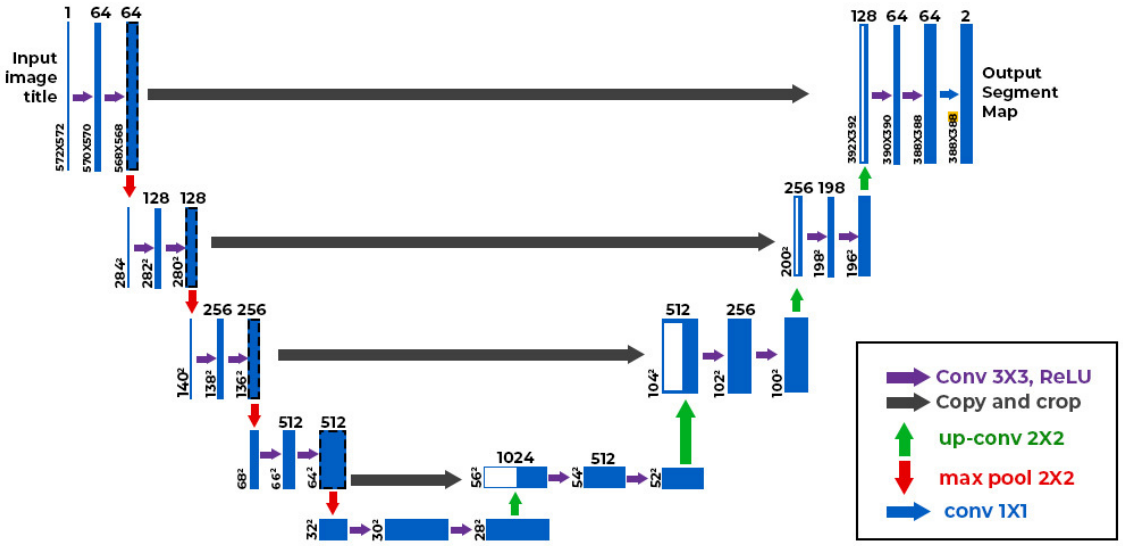

Figure 2.1: Basic U-Net architecture. The blue box corresponds to a multi-channel feature map. The value on top of the box means the number of channels. White boxes denotes he copied feature maps.

2.2 3D U-Net

The 3D U-Net paper [10] published in 2016 is a extension of the original 2D unet paper which focuses on the network for volumetric segmentation which can correctly learn form sparse annotated volumetric images. Ozgun et al. [10] suggested a network which requires 2D slices for training but can generate dense volumetric segmentation. This segmentation is suitable for both semi-automated and fully automated setups. The extensions are from 2D U-Net architecture to 3D with 3D operations like convolutions, max pooling, and up-convolutional layers and includes batch normalization for faster convergence. For instance, the analysis path features layers with $3 \times 3 \times 3$ convolutions followed by ReLU activations and $2 \times 2 \times 2$ max-pooling operations. Conversely, the synthesis path includes $2 \times 2 \times 2$ up-convolutions with strides of two in each dimension, followed by two consecutive $3 \times 3 \times 3$ convolutions and ReLU activations. The network is trained end-to-end from scratch, using elastic deformations for data augmentation, and a weighted loss function to learn from sparse annotations.

2.3 U-Net++

The development of segmentation architectures has seen significant advancements, beginning with Fully Convolutional Networks (FCNs) introduced by Long et al [29]. FCNs laid the groundwork for semantic segmentation by using skip connections to sum up-sampled feature maps with feature maps skipped from the encoder, effectively

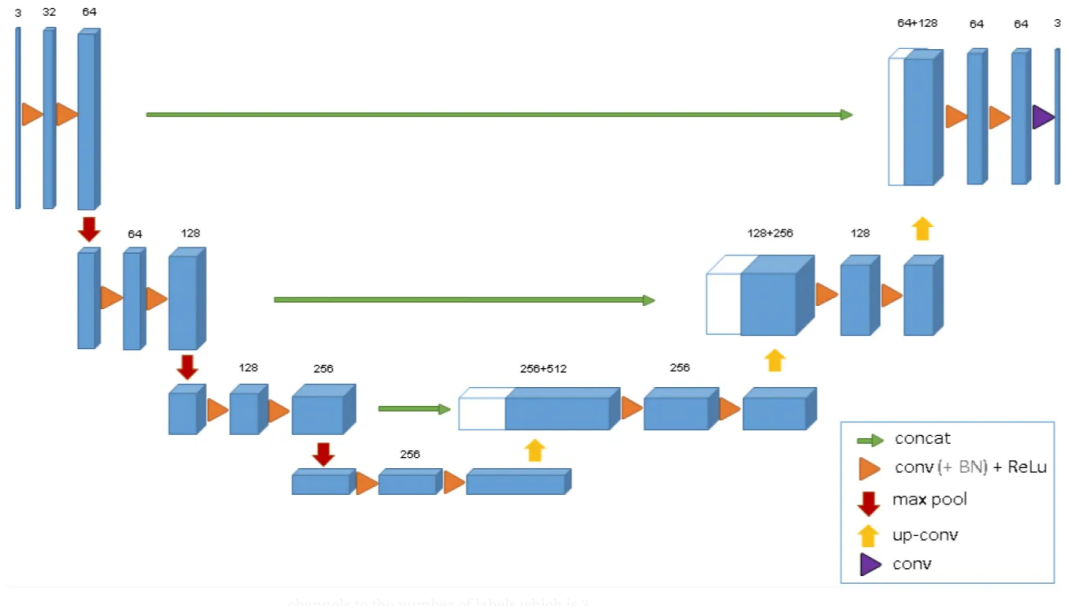

Figure 2.2: 3D U-Net architecture where the boxes are feature maps. The red arrow means Conv3D 3*3*3 with stride 2. The yellow arrow denotes Conv3D Transpose.

recovering full spatial resolution at the network output. This foundational work paved the way for subsequent architectures such as U-Net.

Proposed by Ronneberger et al. [35], U-Net expanded on FCNs by concatenating feature maps from the encoder to the decoder, incorporating convolutions and non-linearities between each up-sampling step. U-Net's effectiveness in recovering spatial information made it widely adopted for biomedical image segmentation. Further advancements were inspired by the DenseNet architecture introduced by Huang et al. [22] and Li et al. [26] developed H-DenseUNet for liver and liver tumor segmentation, integrating dense connections to enhance the flow of information and gradients throughout the network, thus improving segmentation performance.

The critical role of skip connections in biomedical image segmentation was systematically investigated by Drozdzal et al. [17], highlighting their importance in enabling better information flow and improving segmentation accuracy. Concurrently, architectures like GridNet and Mask-RCNN also contributed to the evolution of segmentation networks. GridNet, introduced by Fourure et al. [19], generalized several classical segmentation architectures by wiring feature maps in a grid fashion, though it lacked up-sampling layers between skip connections, setting it apart from U-Net.

Mask-RCNN emerged as a significant meta-framework for object detection, classification, and segmentation. Its flexibility allows for the integration of other architectures, potentially incorporating UNet++ as its backbone by replacing plain skip connections

with nested dense skip pathways.

UNet++ represents a notable modification of the U-Net architecture, redesigning the skip pathways and integrating dense convolution blocks to reduce the semantic gap between encoder and decoder feature maps. This innovative approach enhances segmentation performance, particularly evident in various medical imaging datasets where UNet++ demonstrated significant improvements over U-Net and wide U-Net. The deep supervision mechanism and redesigned skip pathways were crucial in achieving these enhancements, showcasing the incremental advancements that have shaped the field of segmentation architectures. UNet++ aggregates feature maps at each unit follows the equation:

$$x_{i,j} = \begin{cases} H(x_{i-1,j}), & \text{if } j = 0 \\ H\left(\left[x_{i,k}\right]_{k=0}^{j-1}, U(x_{i+1,j-1})\right), & \text{if } j > 0 \end{cases} \tag{2.1}$$

In the equation for the base case ($j = 0$), the feature map $x_{i,j}$ is obtained by applying a convolutional operation to the previous depth level $x_{i-1,j}$. In the recursive case ($j > 0$), $x_{i,j}$ is derived by combining concatenated feature maps from previous blocks with upsampled features from the decoder path, all processed through a convolutional operation. This approach effectively integrates multi-scale information and dense skip connections, leading to enhanced segmentation performance.

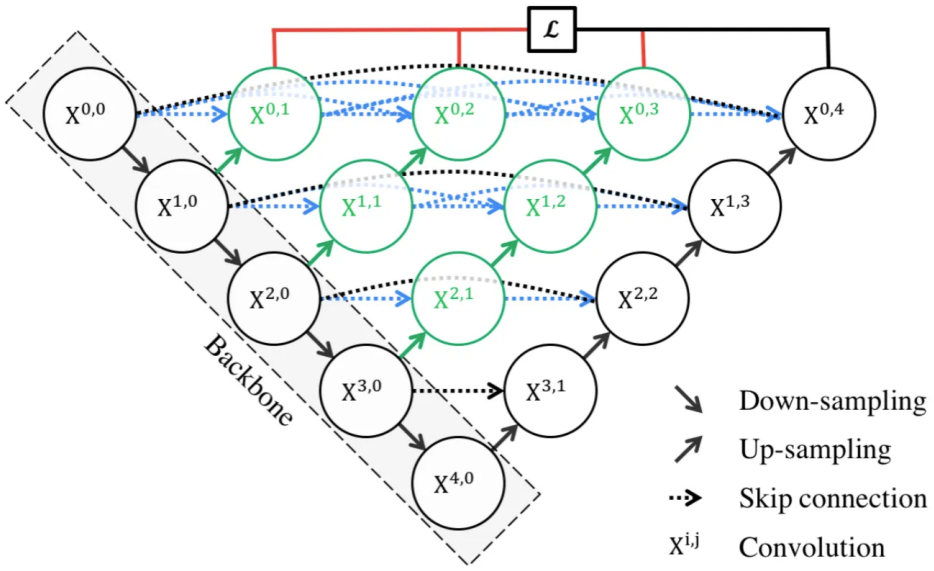

Figure 2.3: Basic U-Net++ architecture

Overall, the evolution from FCNs to UNet++ underscores the continuous efforts to

optimize segmentation accuracy and efficiency, with each new architecture building on the strengths and addressing the limitations of its predecessors.

2.4 Swin U-Net

The Swin-Unet paper [6] provides an insightful overview of the advancements in medical image segmentation, emphasizing the evolution from traditional methods to the latest transformer-based approaches. Initially, medical image segmentation relied heavily on contour-based techniques and traditional machine learning algorithms. These methods, while foundational, were limited in their ability to handle complex segmentation tasks. The introduction of U-Net marked a significant advancement, utilizing a U-shaped architecture with skip connections that greatly improved segmentation performance.

Building on the success of U-Net, several variants have been developed to further enhance segmentation accuracy. Res-UNet [13] incorporated residual connections to address the vanishing gradient problem, while Dense-UNet[5] integrated dense connections to improve feature reuse and gradient flow. UNet++ [48], UNet3+ [23], and 3D-Unet [10] extended the original architecture to better capture fine details and handle three-dimensional medical images. V-Net [1] specifically adapted U-Net for volumetric data, further pushing the boundaries of medical image segmentation. Despite these advancements, CNN-based methods have inherent limitations in capturing global and long-range semantic information due to the localized nature of convolution operations.

The emergence of vision transformers introduced a new paradigm in image analysis. Originally designed for machine translation, transformers achieved state-of-the-art performance in various natural language processing tasks. The Vision Transformer (ViT) [16] adapted these models for vision tasks, demonstrating impressive results in image recognition. DeiT [41] further optimized ViT for mid-size datasets and improved robustness through distillation methods. The Swin Transformer [27], a hierarchical vision transformer using shifted windows, set new benchmarks in image classification, object detection, and semantic segmentation by effectively modeling global and long-range interactions.

Recognizing the limitations of traditional CNNs in capturing long-range dependencies, the Swin-Unet [6] leverages the capabilities of the Swin Transformer. This pure transformer-based U-shaped architecture is designed to enhance segmentation accuracy in medical imaging. By combining the hierarchical design of the Swin Trans-

former with the U-shaped architecture's strengths, Swin-Unet aims to provide superior performance in medical image segmentation, addressing the shortcomings of previous CNN-based methods and setting a new standard in the field.

2.5 UNetR

UNETR, also known as UNET TRansformers [21], represents a significant advancement in the field of medical image segmentation, particularly in handling 3D volumetric data. Traditional FCNN-based architectures, such as U-Net and its variants, have been highly effective in segmentation tasks due to their U-shaped encoder-decoder design. However, these architectures often struggle to capture long-range dependencies because of the localized nature of convolutional layers.

Inspired by the success of transformers in NLP for sequence learning, researchers have explored their potential in medical imaging. Transformers, with their self-attention mechanism, are adept at modeling long-range dependencies and global context. Vision Transformers (ViTs) have demonstrated impressive results in computer vision by treating images as sequences of patches, thus capturing detailed feature representations across different scales.

UNETR leverages these capabilities by utilizing a transformer as the encoder for 3D medical image segmentation. The architecture reformulates the segmentation task as a sequence-to-sequence prediction problem. It directly uses volumetric data, embedding patches for the transformer encoder, which then connects to a CNN-based decoder via skip connections. This combination allows UNETR to maintain the effective U-shaped design while enhancing its ability to learn from the entire volume, improving segmentation accuracy.

Several studies have validated the effectiveness of transformer-based architectures. For instance, Xie et al. [44] proposed a hybrid framework that combines a CNN encoder with a transformer for feature processing and a CNN decoder for output prediction. Similarly, Wang et al. [43] integrated a transformer in the bottleneck of a CNN-based encoder-decoder for brain tumor segmentation. These approaches underscore the potential of transformers to enhance segmentation by capturing detailed spatial relationships and contextual information.

UNETR's unique approach of integrating a transformer encoder with a CNN-based decoder has demonstrated state-of-the-art performance on datasets like BTCV and MSD, particularly in brain tumor and spleen segmentation tasks. This novel architec-

ture addresses the limitations of traditional FCNNs, providing a robust solution for complex medical image segmentation.

In summary, UNETR exemplifies the integration of transformer-based models in medical imaging, addressing the challenges of long-range dependency modeling and enhancing segmentation accuracy. This architecture represents a significant leap forward, paving the way for more advanced and reliable diagnostic tools in the medical field.

2.6 Data Augmentation for 3D brain Tumor Segmentation

In their study, Cirillo et al. [11] explore the efficacy of various data augmentation techniques for 3D brain tumor segmentation using a standard 3D U-Net architecture. This is particularly relevant given the challenges in obtaining large annotated datasets in medical imaging, which often result from stringent data protection regulations and ethical considerations. They emphasize that while data augmentation is a well-known strategy in general machine learning applications, its specific impact on 3D brain tumor segmentation has not been thoroughly investigated. The study highlights that despite the abundance of research in brain tumor segmentation, detailed examinations of augmentation techniques remain sparse. Cirillo et al. [11] compared several augmentation techniques including flipping, rotation, scaling, brightness adjustment, and elastic deformation. They found that brightness augmentation and elastic deformation significantly enhanced the performance of the segmentation network. Specifically, elastic deformation with a standard deviation of 2 and brightness adjustments yielded the best results, with both techniques substantially improving Dice scores across different tumor classes. This suggests that these augmentations effectively simulate the variability seen in real-world medical imaging, thereby enhancing the robustness of the trained models. Other augmentation techniques such as random scaling and rotation also showed positive impacts but were less effective than brightness and elastic deformation. Interestingly, the study found that combining multiple augmentation techniques did not lead to further improvements compared to using single techniques. This outcome indicates that while multiple augmentations could theoretically provide more diverse training samples, the practical benefit might be limited by the proportion of original images presented during training. Furthermore, the paper employed a robust evaluation methodology using 3-fold cross-validation and ensemble averaging to ensure the reliability of their findings. They also performed

non parametric permutation tests to statistically validate the improvements seen with augmentation techniques. The study concluded with practical recommendations for future researchers, suggesting the use of elastic deformation and brightness adjustment as primary augmentation strategies when working with similar datasets. This comprehensive analysis not only underscores the importance of data augmentation in enhancing model performance for 3D brain tumor segmentation but also provides a nuanced understanding of how different augmentation strategies can be applied effectively. The authors' contribution is particularly valuable for advancing the application of deep learning in medical imaging, offering a detailed road-map for optimizing training processes in the face of limited annotated data.

2.7 XAI in Medical Image Analysis

The literature on Explainable Artificial Intelligence (XAI) in the context of medical image analysis emphasizes the need for transparency and interpretability in AI systems, particularly in high-stakes fields such as healthcare. XAI techniques can be broadly categorized into model-based and post hoc explanations. Model-based explanations are integrated into the model's architecture and training process, ensuring that the model is interpretable by design. Conversely, post hoc explanations are applied after the model has been trained, offering insights into the model's decisions without altering its internal workings [42].

Model-specific explanations are tailored to particular types of models, such as convolutional neural networks (CNNs), and leverage the model's structure to provide insights. In contrast, model-agnostic explanations can be applied to any model, offering flexibility but potentially less specificity. The scope of explanations can also vary, with global explanations providing insights into the overall model behavior, and local explanations focusing on individual predictions.

In medical image analysis, visual explanation techniques are prevalent. Backpropagation-based approaches, such as saliency maps and Grad-CAM, highlight important regions in an image that influence the model's predictions. Perturbation-based methods, including occlusion sensitivity and Local Interpretable Model-agnostic Explanations (LIME) [33], assess the impact of modifying specific parts of the input on the output. Textual explanations, such as image captioning and Testing with Concept Activation Vectors (TCAV) [25], provide narrative descriptions of the model's decisions, enhancing interpretability for clinical users.

Example-based explanations, such as prototype networks and influence functions,

use representative cases to elucidate model behavior. These methods are particularly useful in medical contexts, where clinicians often rely on examples and analogies. However, the validity and robustness of these techniques are critical concerns. Evaluations of XAI methods must consider their reliability and faithfulness to the original model's decision-making process.

Critiques of XAI often highlight issues related to the interpretability and practical utility of the explanations. For instance, saliency maps may not always provide clear or meaningful insights, and the fidelity of explanations can vary, potentially leading to misinterpretations. Despite these challenges, the integration of XAI in medical imaging holds promise for improving diagnostic accuracy and fostering trust in AI systems. Future directions for XAI include combining multiple forms of explanation, such as textual and visual, to provide more comprehensive insights and exploring biological explanations to link imaging features with underlying biological processes.

This literature review underscores the ongoing efforts to refine and validate XAI techniques, ensuring they meet the stringent requirements of medical applications and enhance the interpretability of AI-driven diagnostics.

Chapter 3

Materials and Methods

3.1 Overview

The research proposes a dual-faceted approach to brain tumor segmentation utilizing multi-modal MRI data, encompassing both two-dimensional and three-dimensional segmentation models. This chapter offers a comprehensive description of the methodologies employed, encompassing dataset description, Loss functions, performance evaluation metrics, and the incorporation of Explainable Artificial Intelligence (XAI) techniques. Figure 3.1 provides a clear and comprehensive visual depiction of the entire workflow.

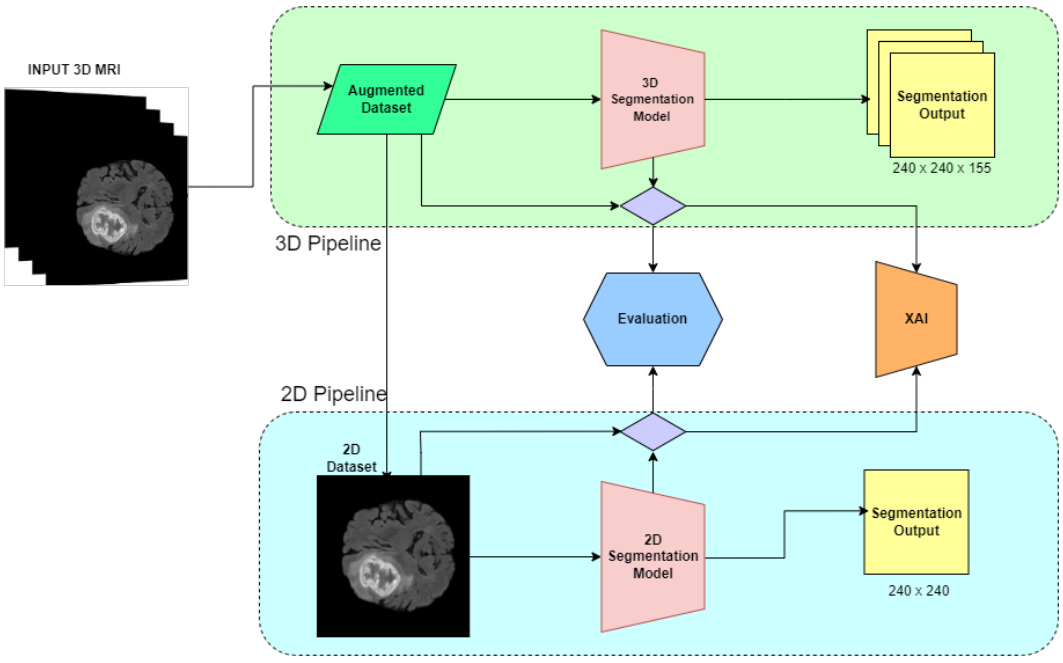

Figure 3.1: Workflow of the proposed 2D and 3D segmentation methodologies.

The workflow presented in Figure 3.1 illustrates a comparative approach for 2D and 3D segmentation methodologies in medical image analysis, specifically focused on MRI data. The process begins with the input of 3D MRI data, which serves as the foundation for both the 3D and 2D pipelines.
In the 3D pipeline, the input data undergoes augmentation to create an expanded dataset. This augmented dataset is then fed into a 3D segmentation model, which processes the volumetric data and produces a segmentation output with dimensions of $240 \times 240 \times 155$.
Parallel to this, the 2D pipeline operates on a slice of the input data. A 2D dataset is extracted from the original 3D input and passed through a 2D segmentation model. This model generates a segmentation output with dimensions of 240×240, corresponding to a single slice of the 3D volume.
Both pipelines converge at an evaluation stage, where the performance and results of the 2D and 3D segmentation approaches can be compared and analyzed. Additionally, both pipelines feed directly into a XAI (Explainable AI) component, suggesting that the workflow incorporates methods for interpreting and understanding the decisions made by the segmentation models.

3.2 Dataset

The BraTS 2020 [31], [3], [4] dataset presents a comprehensive collection of multimodal MRI scans, available in NIfTI format (.nii.gz). This collection encompasses four distinct imaging modalities:

1. Native (T1)
2. Post-contrast T1-weighted (T1Gd)
3. T2-weighted (T2)
4. T2 Fluid Attenuated Inversion Recovery (T2-FLAIR)

These volumetric data were obtained through varying clinical protocols and imaging systems across several participating institutions (n=19). Figure 3.2 presents visual illustrations of these MRI modalities.

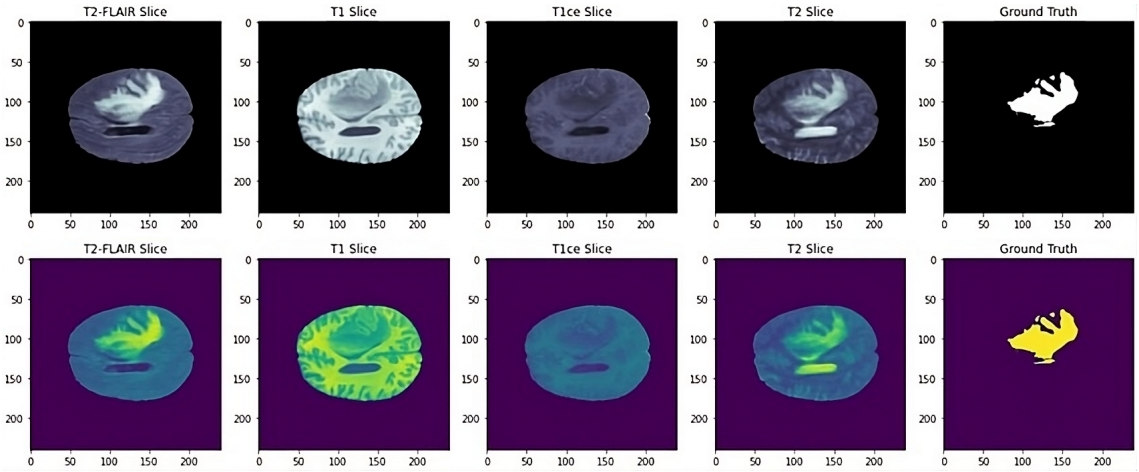

Figure 3.2: Examples of multi-modal MRI scans from the BraTS 2020 dataset: (a) T2-FLAIR, (b) T1, (c) T1ce, (d) T2 and (e) Ground Truth. [45]

The dataset's integrity is assured through a rigorous manual segmentation process, comprising one to four expert raters following to a standardized annotation protocol. Subsequently, these annotations underwent review and validation by qualified neuro-radiologists to guarantee correctness and consistency.

The annotations indicate three major locations of interest:

1. GD-enhancing tumor
2. Peritumoral edema
3. Necrotic and non-enhancing tumor core

An illustrative example of these annotated regions on an MRI scan is provided in Figure 3.3.

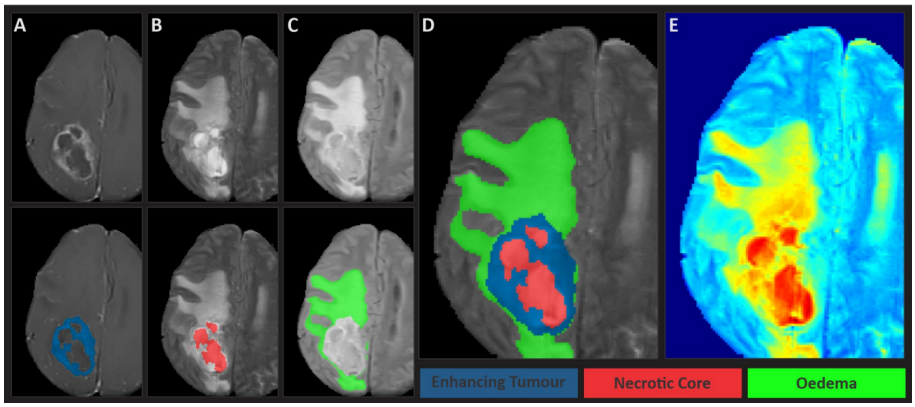

Figure 3.3: Annotated regions in a sample MRI scan: (A) GD-enhancing tumor, (B) necrotic and non-enhancing tumor core, (C) peritumoral edema, (D)Whole tumor region (E) Whole brain scan heatmap .

This meticulous segmentation process boosts the dataset's robustness and reliability, rendering it particularly ideal for the training and evaluation of deep learning models. The integration of different MRI modalities for brain tumor segmentation offers a rich, multi-dimensional information source, facilitating the construction of models capable of exact detection and delineation of diverse tumor regions.

These rigorously curated and annotated datasets play a vital role in the evolution of explainable AI models, hence providing transparency and reliability in brain tumor segmentation tasks.

3.3 Loss Functions

The scope of optimization techniques extends beyond just modeling the architecture. These methods additionally incorporate loss functions and activation functions. Loss functions can be categorized into many sorts, such as those that rely on distribution, region, border, and compound techniques. In this section, we provide a concise overview of several frequently employed loss functions in the field of medical image segmentation.

3.3.1 Cross-Entropy Loss

Here, we are briefly demonstrating some commonly used loss functions used for medical image segmentation. The most used loss function is cross-entropy loss [47], and it is derived from Kullback–Leibler (KL) [12] divergence to evaluate the variation (dissimilarity) between two distributions. It is given by:

$$L_{ce} = -\sum_{i=1}^{N} [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)] \tag{3.1}$$

where p_i refers to the training result, g_i refers to the ground truth, and N is the number of pixels. Cross-entropy loss converges quickly because the gradient of the last layer is not relevant to the activation function, where the difference is only related to the result and the ground truth. Many researchers use cross-entropy loss, but using this loss is preferable when the segmented target is not extremely different from the background. However, region-based loss is more likely to be used when this type of loss aims to minimize the mismatch or maximize the overlapping between the segmentation results and the ground truth.

3.3.2 Dice-loss function

Another widely used loss function is the Dice loss function, used for medical image segmentation. It is extracted from the Sorensen–Dice coefficient [14], and it directly optimizes the mostly used metric for segmentation, which is the Dice coefficient. The Dice loss is given by:

$$D = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \quad \in [0,1] \tag{3.2}$$

Here, g_i is the ground truth pixels (voxels if 3D segmentation task) and N is the number of pixels. Since, in image segmentation networks, the last layer, which is mostly a SoftMax layer, the output is a probability of each pixel belonging to foreground or background. In addition, the Dice loss can be differentiated to produce the gradient:

$$\frac{\partial D}{\partial p_j} = \frac{2|GT \cap Pr|}{|GT| + |Pr|} = 2\left[\frac{g_i\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right) - 2p_i\left(\sum_i^N p_i g_i\right)}{\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right)^2}\right] \tag{3.3}$$

Here, GT is the ground truth. P_r is the output segmentation result. The aim of Dice loss is to establish the right balance between the foreground (target) and the background by tuning the weight matrices.
Another extension of Dice loss is the generalized Wasserstein Dice loss used for multi-class segmentation, which takes the advantages of the hierarchical structure of complicated tissues.

3.3.3 IoU Loss

Intersection over union [18], or Jaccard loss, is identical to Dice loss and belongs to the same category of region-based loss. It is derived from the Jaccard index, and it simply measures the intersection between the segmentation results and the ground truth. It is given by:

$$IoU = \frac{|GT \cap Pr|}{|GT \cup Pr|} = \frac{|GT \cap Pr|}{|GT| + |Pr| - |GT \cap Pr|} = \frac{\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \epsilon\, [0,1] \tag{3.4}$$

Here, p_i refers to the training result, g_i refers to the ground truth, GT is the ground truth. P_r is the output segmentation result.

3.3.4 Tversky Loss

This loss [36] is also a region-based loss and is a modified Dice loss. It sets different weights to the false negative (FN) and false positive (FP), whereas Dice loss uses the same weights for the preceding terms. This makes Tversky loss suitable for the unbalanced datasets. The Tversky loss formula is given by:

$$T_I = \frac{|GT \cap Pr|}{|GT \cap Pr| + \alpha|Pr \setminus GT| + \beta|GT \setminus Pr|} \tag{3.5}$$

It is also formulated as:

$$T(\alpha, \beta) = \frac{\sum_{i=1}^{N} p_{ic} g_{ic}}{\sum_{i=1}^{N} p_{ic} g_{ic} + \alpha \sum_{i=1}^{N} p_{i\bar{c}} g_{ic} + \beta \sum_{i=1}^{N} p_{ic} g_{i\bar{c}}} \tag{3.6}$$

where p_{ic} is the probability that pixel i is from class c, $p_{i\bar{c}}$ is the probability that pixel c is not from the class c (for example, class c means tumor tissue), and same terminology applies for g_{ic} and $g_{i\bar{c}}$ considering it is ground truth pixels. α and β are the hyperparameters, and tuning these two parameters can shift the emphasis to better the recall when having class imbalance .

3.3.5 Hausdroff Distance Loss

Finally, the boundary losses category aims to minimize the distance between both the segmentation result and the ground truth . It is used for extremely unbalanced data, and the most used boundary loss function is Hausdorff distance loss [24], which tries to estimate the Hausdorff distance from the network output probability and aims to reduce it. Hausdorff distance loss is given by:

$$L_{HD} = \frac{1}{N} \sum_{i=1}^{N} \left[(p_i - g_i) \circ \left(d_{Gi}^2 + d_{pi}^2 \right) \right] \tag{3.7}$$

where p_i refers to the training result, g_i refers to the ground truth, d_{Gi} and d_{pi} are the distances of the ground truth and segmented result, respectively and $\circ$ is the Hadamard Product (entry-wise).

Lastly, compound loss functions are used by summing over various types of loss functions to produce new mixed loss functions.

3.4 Evaluation Metrics

Choosing the right metric for evaluating any task in deep learning is vital because specific metrics are used to evaluate different tasks. In this section, we will briefly present the widely used metrics for medical image segmentation. Rather than elaborating about the evaluation metrics used for different tasks that use deep learning for medical image analysis, we will only focus on the metrics for segmentation.

3.4.1 Dice Coefficient

The first and the most common metric for validating medical volume segmentation is called Dice-score coefficient (DSC) (or overlap index) [14]. This metric is defined by Equation (24):

$$DICE = \frac{2|S_g \cap S_p|}{|S_g| + |S_p|} = \frac{2TP}{2TP + FP + FN}, \epsilon\, [0, 1] \tag{3.8}$$

S_g and S_p are the segmented region of ground truth and the predicted segmentation result, respectively. The value of DICE is "0" when there is no overlapping between the resulting segmented area and the ground truth segmented area, and it is equal to "1" when they are 100% overlapped. Since the confusion matrix calculates all the distribution probabilities, many evaluation metrics are derived from the terminologies corresponding to the confusion matrix, such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

3.4.2 Jaccard Index/Intersection over Union (IoU)

IoU [18] calculates overlapping area between the ground truth and the segmentation result divided by their union. Therefore, it gives an idea about the similarity between both regions. It is given by the formula:

$$JAC = IoU = \frac{|S_g \cap S_p|}{|S_g| + |S_p|} = \frac{TP}{TP + FP + FN}, \epsilon\, [0, 1] \tag{3.9}$$

From the equation above, we note that the difference between DICE and IoU is that IoU is always greater than DICE, except at the peak {0,1} where they are equal. In addition, the relation between both metrics is given by:

$$JAC = IoU = \frac{|S_g \cap S_p|}{|S_g| + |S_p|} = \frac{2|S_g \cap S_p|}{2(|S_g| + |S_p| - |S_g \cap S_p|)} = \frac{DICE}{2 - DICE} \tag{3.10}$$

Similarly

$$DICE = \frac{2JAC}{1 + JAC} \tag{3.11}$$

Which means that both metrics measure the same aspects and evaluate the system ranking, hence, selecting one metric to evaluate the results is enough.

3.4.3 Hausdorff Distance (HD)

It is one of the recent rising used metrics for evaluation of a segmentation task, however, reducing the Hausdorff distance [24] is the goal of segmentation because it is evidence of the segmentation error. For two-point sets, X and Y, the distance from X to Y is defined as:

$$HD(X, Y) = \frac{1}{N} \sum_{x \in X} \min_{y \in Y} ||x - y|| \tag{3.12}$$

where N is the total number of observations (voxels or pixels).

Moreover, the average Hausdorff distance between X and Y is given by:

$$d_{AHD}(X, Y) = \left(\frac{1}{X} \sum_{x \in X} \min_{y \in Y} d(x, y) + \frac{1}{Y} \sum_{y \in Y} \min_{x \in X} d(x, y) \right) / 2 \tag{3.13}$$

Here, the average Hausdorff distance can be calculated as the mean of the directed average from X to Y and from Y to X.

For the medical image segmentation, we assume that point set X, and point set Y are the ground truth voxels and the segmentation result voxels, respectively. Therefore, the HD can be calculated in millimeters or voxels, then Equation (29) can be written as:

$$HD_{avg} = \left(\frac{G \text{ to } S}{G} + \frac{S \text{ to } G}{S} \right) / 2 \tag{3.14}$$

where G to S is the directed average HD from the ground truth to the segmentation

result, and vice versa for the term S to G, where G and S are the voxels of the ground truth and the segmentation result, respectively. HD is sensitive to outliers.

3.4.4 Sensetivity and Specificity

Also called true positive rate (TPR) or recall, this metric measures the positive pixels fraction in the ground truth, which also are predicted as positive in the segmented result. Similarly, true negative rate (TNR) or specificity gauges the negative pixels (background) that are identified as negative pixels from the ground truth and the segmentation result. These two metrics are both valuable because of their sensitivity to the segment sizes, which make them suitable for segmenting small size regions (e.g., retina vessels) because they penalize the small segments [20]. We demonstrate the formula of sensitivity, and specificity as:

$$Recall = Sensitivity = TPR = \frac{TP}{TP + FN} \tag{3.15}$$

$$Specificity = TNR = \frac{TN}{TN + FP} \tag{3.16}$$

The preceding metrics are the major metrics used for medical image segmentation, and there are other evaluation metrics, but they less common, which are highlighted in [32], [40].

3.5 Explainable AI

The model was analysed using grad-cam, feature ablation and kernal shap. The output of these xai methods are heatmaps which tells us which region of the image is the model focusing on to predict a certain class. In our case there are four classes. Whole Tumor, Tumor Core, Enhancing Tumor and Background.

Output of these XAI methods can be understood by the color of the pixels. Red indicates that the pixel contributes negatively, and green indicates that the pixel has positive contribution to predicting a certain class. So these heatmaps helps us visualize which part of the image is important for predicting a certain class and which part are irrelevant (blue regions).

3.5.1 GradCAM

Grad-CAM is a popular technique in explainable AI (XAI) for providing visual explanations of deep learning models, particularly convolutional neural networks (CNNs). It highlights the regions of an input image that are most influential in the model's decision-making process by using the gradients of the target class flowing into the final convolutional layer [38].

These are the steps used in this method:

1. Forward Pass: Pass the input image X through the CNN to obtain the output feature maps of the final convolutional layer. Let A^k denote the k-th feature map of this layer, and Y represent the ground truth labels or the predicted output.

2. Class Score: Compute the score for the target class c. This is usually the output before the softmax layer in classification models, denoted as y^c.

3. Gradient Calculation: Compute the gradients of the score for class c with respect to the feature maps A^k:

$$\frac{\partial y^c}{\partial A^k} \tag{3.17}$$

   These gradients indicate how much the target class score changes with respect to a small change in the feature maps.

4. Global Average Pooling: Perform global average pooling on these gradients to obtain the importance weights α_k^c for each feature map k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.18}$$

   where Z is the number of pixels in the feature map (i.e., the product of the height and width of the feature map), and A_{ij}^k denotes the value at position (i, j) in the k-th feature map.

5. Weighted Combination: Combine the feature maps A^k using the importance weights α_k^c to obtain the class activation map $L_{\text{Grad-CAM}}^c$:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{3.19}$$

   The ReLU function ensures that only the features that have a positive influence on the class score are considered, which makes the map more interpretable by focusing on the regions that contribute positively to the prediction.

6. Upsample to Input Image Size: The resulting class activation map $L^c_{\text{Grad-CAM}}$ is typically smaller than the input image. Upsample $L^c_{\text{Grad-CAM}}$ to the size of the input image using bilinear interpolation.

3.5.2 Feature Ablation

Feature ablation is a technique used in explainable AI (XAI) to understand the importance of specific features or regions in the input image by systematically removing them and observing the impact on the model's performance. Unlike feature occlusion, where features are masked but still present in the input, feature ablation involves completely removing or zeroing out specific parts of the input.

This are the steps taken in this method:

1. Baseline Performance:

   First, compute the model's performance on the original, unablated image. Let X represent the original input image and Y represent the ground truth segmentation mask. The baseline performance P_{baseline} is given by:

   $$P_{\text{baseline}} = f(X, Y) \tag{3.20}$$

   where f is the performance metric, such as accuracy, loss, or mean Intersection over Union (mIoU).

2. Ablation Process:

   Define an ablation mask A of the same dimensions as the input image. The mask will ablate (set to zero or another value) a specific region of the image. Let X_A represent the input image with the ablation applied. For example, $A_{i,j}$ could be a mask that ablates a patch centered at position (i, j).

   The ablated image X_A can be represented as:

   $$X_A = X \odot (1 - A) \tag{3.21}$$

   where $\odot$ denotes element-wise multiplication, and $(1 - A)$ is the inverse mask that leaves all other parts of the image unchanged.

3. Evaluate with Ablated Features

   Compute the model's performance on the ablated image. Let $P_{\text{ablated}}(A)$ denote

the performance with the ablation applied:

$$P_{\text{ablated}}(A) = f(X_A, Y) \tag{3.22}$$

4. Feature Importance:

   The importance of the ablated region is determined by the change in performance caused by the ablation. This change is computed as the difference between the baseline performance and the performance with the ablated region:

$$\Delta P(A) = P_{\text{baseline}} - P_{\text{ablated}}(A) \tag{3.23}$$

   A larger value of $\Delta P(A)$ indicates that the ablated region was more important for the model's predictions.

3.5.2.1 Sliding Window Approach

To systematically ablate different parts of the image, a sliding window approach is often used. This involves moving the ablation mask across the image in a grid-like fashion, ablating different regions, and computing the importance for each region.

1. Define Window Size and Stride: Let the ablation window be of size $w \times h$ and move it across the image with a stride s in both horizontal and vertical directions.

2. Iterate Over Image: For each position (i, j) in the image:

$$X_{A_{i,j}} = X \odot (1 - A_{i,j}) \tag{3.24}$$

where $A_{i,j}$ is the mask centered at (i, j).

3. Compute Performance Impact: For each position, compute the performance impact $\Delta P(A_{i,j})$:

$$\Delta P(A_{i,j}) = P_{\text{baseline}} - P_{\text{ablated}}(A_{i,j}) \tag{3.25}$$

4. Generate Heatmap: Create a heatmap of feature importance where each position (i, j) corresponds to the computed importance $\Delta P(A_{i,j})$.

3.5.3 Kernel SHAP

Kernel SHAP (SHapley Additive exPlanations) is an approach that interprets model predictions by attributing the prediction of an instance to its features based on coop-

erative game theory. SHAP values provide a unified measure of feature importance, ensuring consistency and local accuracy. [30]

3.5.3.1 Shapley Values

Shapley values are calculated for each feature based on the contribution of the feature to the prediction over all possible subsets of features. The formula for Shapley value ϕ_i for feature i is:

$$\phi_i = \sum_{S \subseteq N} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{3.26}$$

In this equation:

- N is the set of all features.
- S is a subset of N that does not include feature i.
- $|S|$ is the number of features in subset S.
- $f(S)$ is the prediction made by the model when only the features in subset S are used.
- $f(S \cup \{i\})$ is the prediction made by the model when the features in subset S and feature i are used.
- The term $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ is a weighting factor that accounts for the different possible orders in which features can be added.

3.5.3.2 Kernel SHAP Approximation

Exact computation of Shapley values is computationally expensive for large models. Kernel SHAP approximates Shapley values using a kernel-weighted linear regression. It samples feature subsets and fits a weighted linear model where the weights are given by the Shapley kernel:

$$\text{Weight}(S) = \frac{(|N| - 1)}{\binom{|N|}{|S|} |S|(|N| - |S|)} \tag{3.27}$$

3.5.3.3 Linear Regression

Solve the weighted linear regression to find the Shapley values for each feature:

$$\hat{y} = \phi_0 + \sum_{i}^{|N|} \phi_i x_i \tag{3.28}$$

where ϕ_0 is the base value (average prediction) and ϕ_i are the estimated Shapley values.

In this equation:

- $\hat{y}$ is the predicted value.
- ϕ_0 is the base value, which is the average prediction over all instances.
- ϕ_i are the estimated Shapley values for each feature i.
- x_i are the values of the features for a specific instance.
- The summation $\sum_{i}^{|N|} \phi_i x_i$ represents the weighted contribution of all features to the prediction.

Chapter 4

Results and Discussion

4.1 3D pipeline

A comprehensive 3D processing pipeline utilized in this study, beginning with data preparation and concluding with the presentation of experimental results. The pipeline leverages the BraTS Challenge 2020 dataset, comprising multi-modal MRI volumes from patients with gliomas. To address the inherent challenges of limited training data, a variety of data augmentation techniques are implemented using the MONAI framework, which enhances the diversity of the training set.

The experimental setup includes details on the hardware and software environment, along with the specific configurations used during model training. The models evaluated include 3D U-Net, U-NetR, and Swin U-NetR, each incorporating different architectural advancements to improve segmentation performance. Finally, the section concludes with a comparative analysis of the models based on the Dice Similarity Coefficient (DSC) across three critical tumor regions: Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT). The results highlight the trade-offs between model complexity and segmentation accuracy, with particular emphasis on the improvements achieved by incorporating residual connections and Swin Transformers into the U-Net architecture.

4.1.1 Data Preparation

The BraTS Challenge 2020 dataset [31], [3], [4] is used for this study, consisting of MRI volumes from patients with gliomas. The dataset includes four types of MRI sequences: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2

Fluid Attenuated Inversion Recovery (FLAIR). All images are preprocessed to have the same resolution and are skull-stripped.

4.1.1.1 Data Augmentation Techniques

To address the challenge of limited training data, various data augmentation techniques are applied to increase the diversity of the training set. The following augmentation methods are implemented using the MONAI framework [7]:

- **Patch Extraction:** From each original volume, a sub-volume of shape 128×128×128 voxels is extracted around its center to ensure the sub-volume predominantly contains brain tissue.
- **Flipping:** Random flipping along one of the three axes (axial, sagittal, and coronal) with a probability of 0.5.
- **Rotation:** Random rotations applied to each axis with angles chosen uniformly within the range [0°, 90°].
- **Scaling:** Scaling factors randomly chosen from a uniform distribution in the range [±10%, ±20%] are applied to each axis.
- **Brightness Adjustment:** Intensity values are adjusted using power-law transformations with the formula $I_{new} = g \cdot I^{\gamma}$. The parameters g and γ are randomly chosen from the range [0.8, 1.2].
- **Elastic Deformation:** Elastic deformation is applied using a square deformation grid with displacements sampled from a normal distribution with standard deviations $\sigma = 2, 5, 8, 10$ voxels.

4.1.2 Experimental Setup

The training dataset was split into 80% for training (295 MRIs) and 20% for validation (74 MRIs), where the 4 modalities were used to generate the 4-channel volume. The labels provided by the dataset (ET, NET-NCR, ED) were converted into 3-channel volume and labeled as enhanced tumor (ET), tumor core (TC), and whole tumor (WT). The generic flow of pre-processing was followed in our experiments using the Medical Open Network for Artificial Intelligence (MONAI) framework[7], where all MRIs are cropped to a smaller size to minimize the computation's need, and the output volume has $128 \times 128 \times 128$ dimensions.

It was found that after epoch 200, the Dice score did not improve, so we considered it

as the main number of epochs.

The experimental setup and configurations are as follows:

- Ubuntu 20.04
- NVIDIA RTX 3090 48GB memory and 128 GB of RAM
- Software used are python 3.9 and cuda 11.3

The main hyperparameters and attributes used are included in Table 4.1

Table 4.1: Hyperparameters used for model training

Hyperparameter	**Value**
Activation Function	Sigmoid
Epochs	200
Loss function	Dice loss
Optimizer	Adam

4.1.3 Experimental Results

We have used the same hyperparameters for all used models. The U-Net architectures used are:

- **3D U-Net**: This model has four levels of convolutions in the encoder and decoder for volumetric data, with skip connections preserving spatial information.
- **U-NetR**: An enhanced U-Net with residual connections within the convolutional blocks, maintaining the four-level encoder-decoder structure. It improves training depth and feature learning.
- **Swin U-NetR**: Combines Swin Transformers with U-NetR. The Swin Transformer replaces convolutional blocks, enhancing long-range dependency capture and contextual understanding in the encoder-decoder framework.

The segmentation results achieved for the 3 models are shown in table 4.2

Table 4.2: Segmentation results on the BraTS 2020 validation dataset for the three experimented models.

Model	DSC			Parameters
	ET	TC	WT	
Unet	0.65	0.653	0.768	48M
UnetR	0.770	0.770	0.843	124M
Swin-UnetR	0.729	0.736	0.807	59.7M

In table 4.2 , three models are compared using Dice Similarity Coefficient (DSC) . DSC is a metric that measures the overlap between the predicted segmentation and the ground truth. The DSC is calculated for three different tumor regions : Enhancing Tumor (ET), Tumor Core (TC) and Whole Tumor (WT).

The UNet has the fewest parameters but shows the lowest DSC scores across all regions, suggesting less effective segmentation capabilities. UNetR significantly improves upon UNet in both DSC scores and, despite its increased parameter count, might offer a better trade-off between complexity and performance. Swin-UNetR shows further improvements in DSC for ET and TC and a slight decrease for WT compared to UNetR but with fewer parameters than UNetR, indicating a more efficient model design possibly due to the incorporation of Swin Transformers.

4.2 2D pipeline

A U-Net model was used for segmenting MRI images using four distinct MRI modalities. Data preparation included resizing the images and normalizing them with MinMaxScaler due to the 216-bit intensity of each channel. The target images underwent one-hot encoding, classifying each pixel into one of four categories: whole tumor (WT), tumor core (TC), enhancing tumor (ET), and background (BG). The model was trained using PyTorch 2.2 on Kaggle, with hyperparameters such as 100 epochs, a batch size of 8, and the Adam optimizer, while loss was calculated with a combination of BCE and Dice loss. The U-Net architecture comprised a contracting path for downsampling and an expansive path for upsampling, utilizing 3x3 convolutions and ReLU activation. Experimental results revealed underfitting, as indicated by the training and validation loss curves, and the Dice and Jaccard scores. Explainable AI methods like GradCAM, feature ablation, and Kernel SHAP were employed to visualize the regions contributing to the model's predictions, highlighting the importance of certain image areas in classification tasks.

4.2.1 Data Preparation

The datasets contain 4 distinct MRI modalities. Both input and output images are resized to the required shape of the U-net model. And lastly, the input and output images are normalized using MinMaxScaler because the intensity of each channel is 216 bits. For the Segmented images, one hot encoding has been used. The true label is divided into 4 channels, whole tumor(WT), tumor core(TC), enhancing tumor(ET) and background(BG). They are the segmentation classes that the model predicts for each pixel.

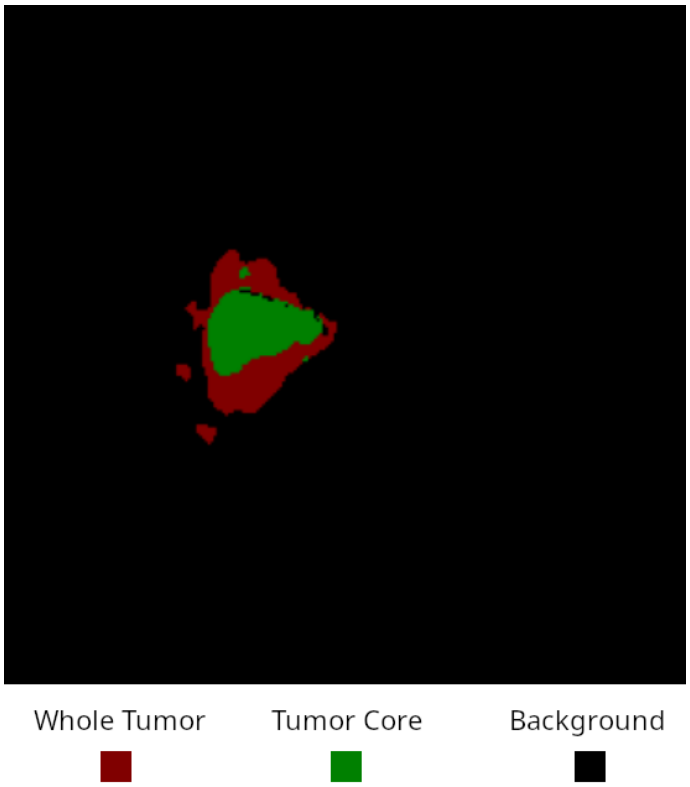

Figure 4.1: Segmantation mask

Figure 4.1 highlights Whole Tumor class as red, Tumor Core class as green and Background as black region.

4.2.2 Experimental Setup

The model implementation was done with pytroch 2.2 and the training was done on kaggle. Table 4.3 outlines the hyperparameters used during the training of the U-Net model. The model was trained for 100 epochs with a batch size of 8. The Adam optimizer was utilized to update the model weights. For loss calculation, a combination of Binary Cross-Entropy (BCE) loss and Dice loss was employed. The performance

of the model was evaluated using the Dice and Jaccard metrics, which are commonly used for segmentation tasks.

Table 4.3: Hyperparameters used for model training

Hyperparameter	Value
Number of Epochs	100
Batch Size	8
Optimizer	Adam
Loss	BCE loss + Dice loss
Metrics	Dice, Jaccard

4.2.3 Network Architecture

Figure 4.2 depicts the 2D U-Net network architecture. The expanding path on the right side of figure 4.2 and the contracting path on the left are the two main paths of it. The typical convolutional network architecture constitutes the contracting path. It consists of two 3x3 convolutions and an activation function, namely, rectified linear unit (ReLU). To downsample, it uses a 2x2 max pooling operation. The number of features doubles with each down-sampling step. 2x2 transposed convo- lutional layers constituted the expansive path, which is used for upsampling the feature map and concatenation with the affiliated feature map of the contracting path, and two 3x3 convolutions with ReLU as an activation function. At each step of upsampling, the number of features is divided by 2. At the last layer, a 1x1 convolution has been used for mapping each feature vector to the wanted number of classes.

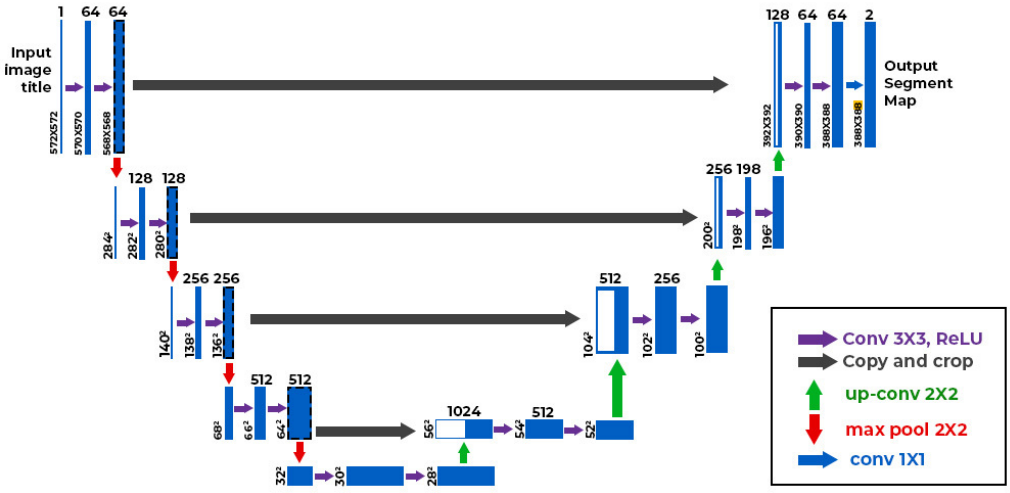

Figure 4.2: Architecture of UNet

4.2.4 Experimental Results

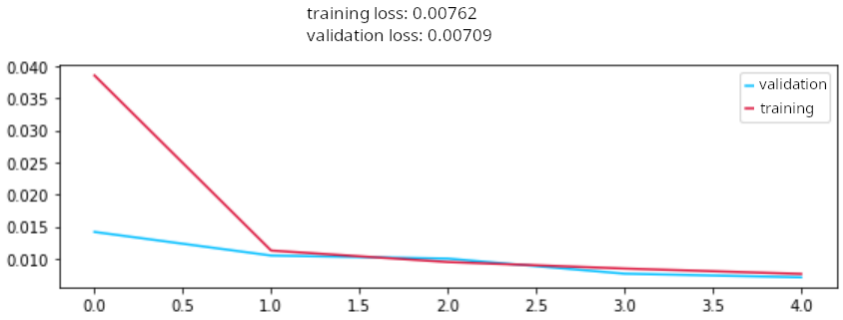

Figure 4.3: Loss curve

From the loss curve in figure 4.3 we can see that, there was a big difference between validation loss and training loss at the start and it is expected. At first the training loss was much higher than the validation loss because of the difference in the number of samples in the traning set and the validation set. In an ideal case, the training loss should have been lower than the validation loss. But at the end, the training loss and validation loss becomes almost thee same. This indicates that there is some under fitting. This is confirmed from dice score in figure 4.4. Similarly as before we can see that the dice score for the training data starts of lower than the validation dice score and at the end, they stabilize the validation dice score crosses the training dice score. We can see the same trend in jaccard score in figure 4.5. In both the curves, the validation score crosses the training score.

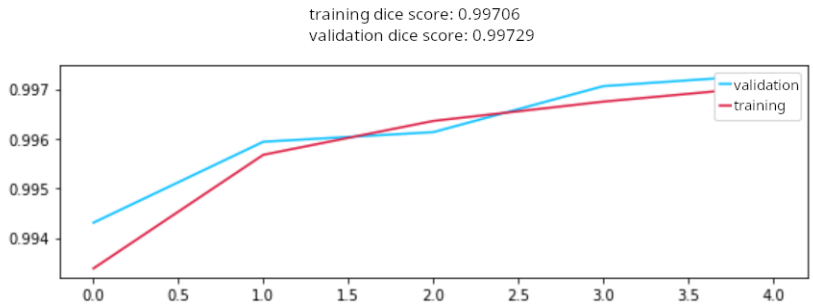

Figure 4.4: Dice score

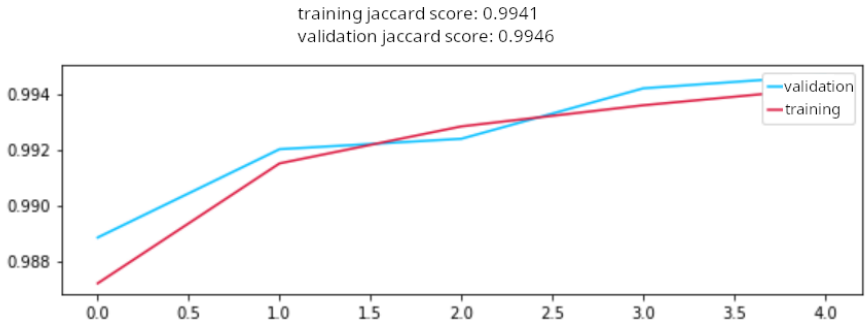

Figure 4.5: Jaccard Score

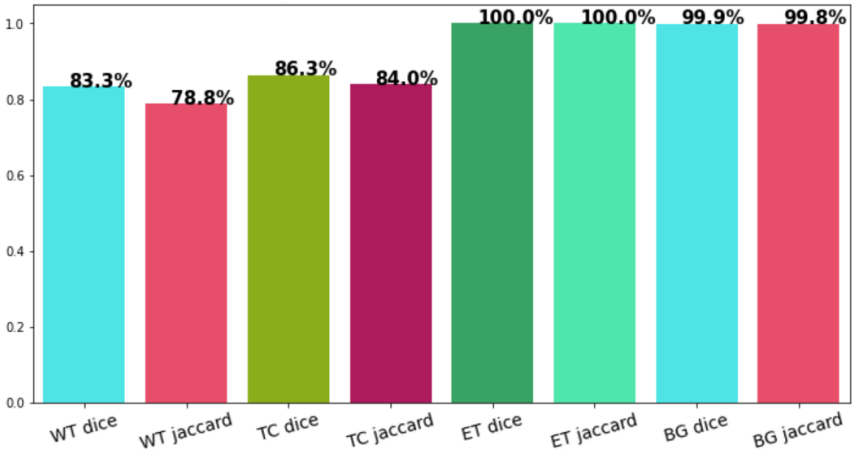

Figure 4.6: Dice and Jaccard coefficient from validation

4.2.5 Explainable AI

The model was analysed using gradcam, feature ablation and kernal shap. The output of these xai methods are heatmaps which tells us which region of the image is the model focusing on to predict a certain class. In our case there are four classes Whole Tumor(WT), Tumore Core(TC), Enhancing Tumor(ET) and Background(BG).

All the xai method output can be understood by the color of the pixels. Red indicates that the pixel contributes negatively, and green indicates that the pixel has positive contribution to predicting a certain class. These heatmaps helps us visualize which part of the image is important for predicting a certain class and which part are irrelevant (blue regions).

4.2.5.1 GradCAM

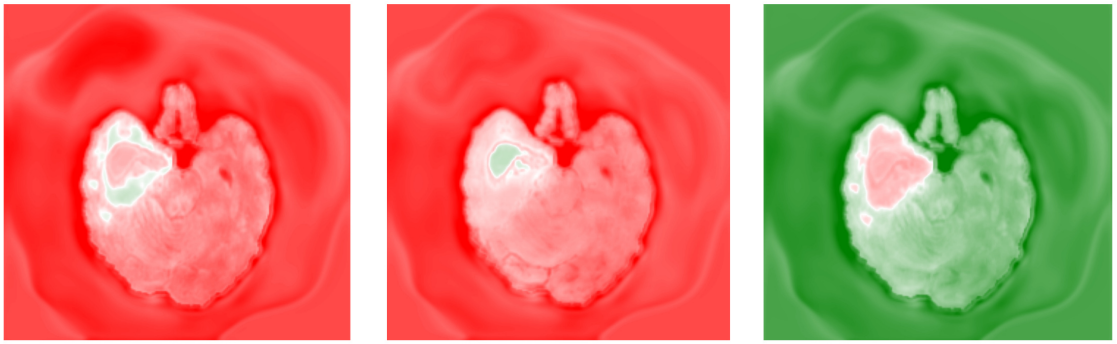

Figure 4.7: Full GradCAM heatmap

Figure 4.7 shows full GradCam heatmap for Whole Tumor class, Tumore Core class and Background class. In the image for the Tumor Core class, the heatmap is mostly red except the region inside the tumor and that is the region where the pixels are classified as Tumor Core. As those regions have higher positive contribution, the heatmap color in that region is greenish instead of bright red like other regions. For the Whole Tumor class, the greenish region is on the boundary of the tumor. So, the boundary of the tumor has positive contributions to predicting Whole Tumor class. And for the Background class, most regions are green except the tumor region which is the expected outcome. This explains that the model is looking the correct regions for predicting Background class and treating the tumor region as something that is very less likely to be classified as Background class and thus indicating high accuracy.

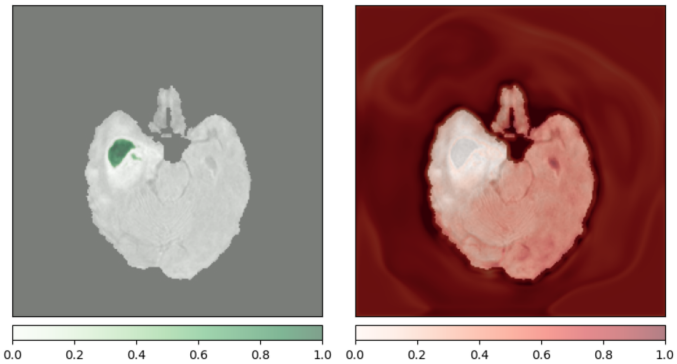

Figure 4.8: GradCAM heatmap for Tumor Core class (postive - negative)

We have also generated the gradcam heatmaps that shows the positive contribution and negative contribution separately.

Figure 4.8 shows two heatmaps. The left one shows the heatmap that contributes only positively and the right side shows the heatmap that contributes only negatively. If

we compare figure 4.8 and the Tumor core image from figure 4.7, we can see that only the greenish region from figure 4.7 is highlighted in the positive contribution image in figure 4.8.

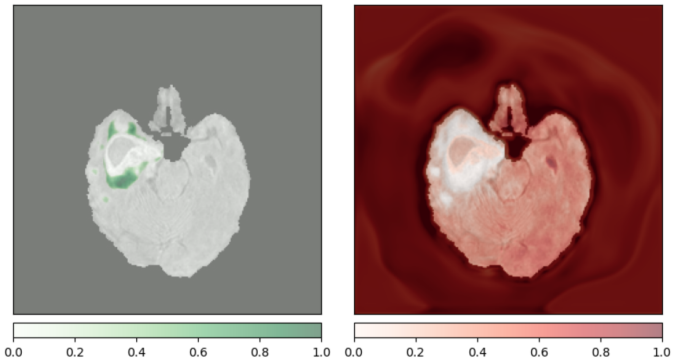

Figure 4.9: GradCAM heatmap for Whole Tumor class (postive - negative)

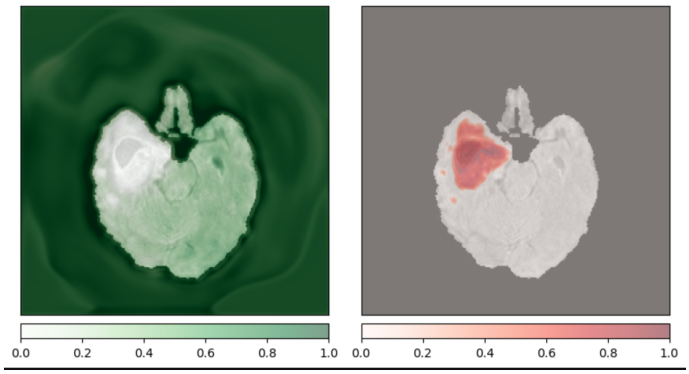

Figure 4.10: GradCAM heatmap for Background class (postive - negative)

4.2.5.2 Feature Ablation

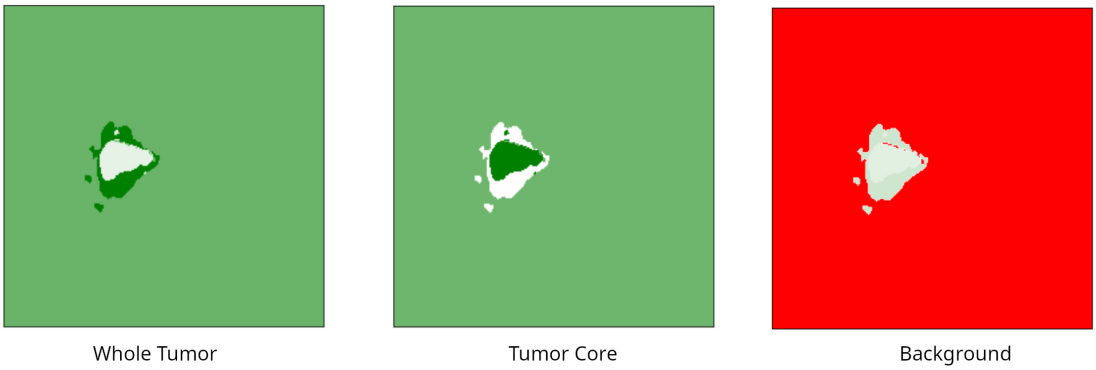

Figure 4.11: Heatmap generated by feature ablation on Tumor Core class

Figure 4.11 shows heatmaps generated by the feature ablation xai technique. The first image shows the heatmap for Whole Tumor class. The dark green region lies on the tumor boundary which is consistent with the gradcam output. For the second image

with Tumor Core class, We can see the inside of the tumor region has dark green color. So those regions are positively contributing to predicting the tumor core class. We can also see that the background is also contributing positively. In the third image with Background class, the background region is bright red indicating that the model is very less likely to predict

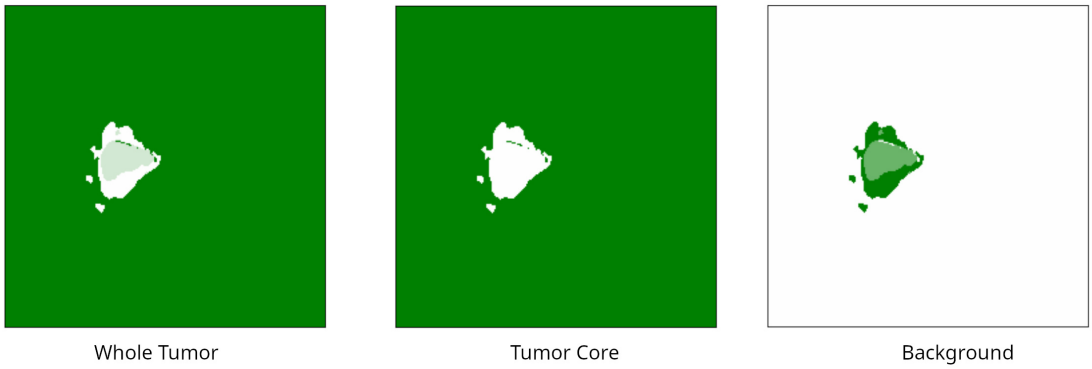

Figure 4.12: Heatmap generated by feature ablation without target class

To investigate further, we have substracted the target class from the heatmap. Figure 4.12 confirms that the backgrounds are indeed used by the model to predict tumor core class. In fact this is true for all classes.

4.2.5.3 Kernel SHAP

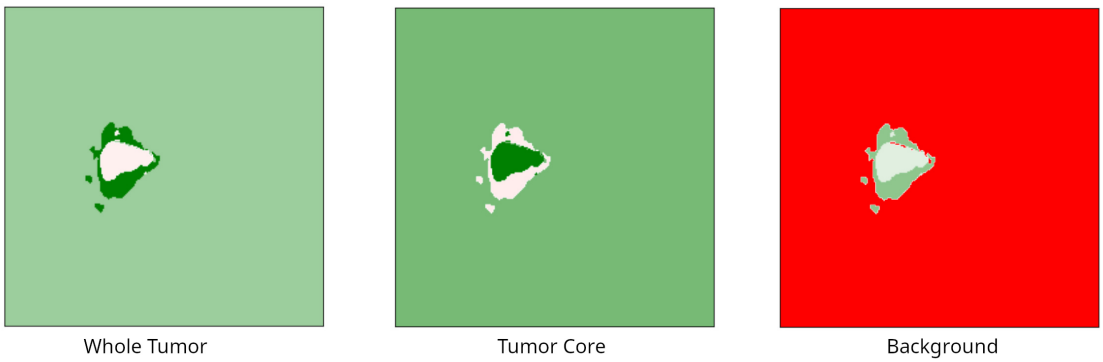

Figure 4.13: Heatmap generated by kernal shap

Figure 4.13 shows the heatmap generated by Kernel SHAP method. We can see the inside of the tumor region has dark green color similar to feature ablation for the image in Tumor Core class. So those regions are positively contributing to predicting the tumor core class. And the boundary of the tumor is of red color indicating that it contributes negatively to predicting tumor core(TC) class. We know that the boundary of the tumor is of whole tumor(WT) class. The negative contribution of that region is helping the model to predict tumor core class.

Similar to feature ablation, we can see in 4.14 the background is contributing positively.

Figure 4.14: Heatmap generated by kernel shap without target class

Chapter 5

Conclusion

In this thesis, we explored the efficacy of various deep learning models—2D U-Net, 3D U-Net, UNETR, and Swin UNETR—for the segmentation of brain tumors in multimodal MRI scans. Our comprehensive analysis revealed that the 2D U-Net model consistently outperformed its 3D counterparts in terms of segmentation accuracy and computational efficiency. Among the 3D models, UNETR emerged as the most effective, demonstrating superior performance in capturing the complex structures of brain tumors.

The comparative evaluation of 2D and 3D models highlighted the strengths and limitations of each approach. While 2D models exhibited better overall performance, the 3D UNETR model showed promise in handling volumetric data, which is crucial for precise tumor segmentation in clinical settings.

To enhance the interpretability of these models, we employed explainable AI techniques, including Grad-CAM and Kernel SHAP, primarily on the 2D U-Net model. These techniques provided valuable insights into how small regions within the MRI scans can significantly impact the segmentation results, highlighting the importance of careful consideration in model interpretation.

However, the study also had limitations, particularly regarding the application of explainable AI techniques, which were only implemented for the 2D models. This constraint limited our ability to fully understand the decision-making processes of the 3D models. Future work should address this gap by extending the application of explainable AI to the 3D models, which would provide a more comprehensive understanding of their inner workings and potentially lead to further improvements in segmentation accuracy.

In summary, this thesis has demonstrated the potential of both 2D and 3D models for brain tumor segmentation, with 2D U-Net leading in performance. However, the promising results from the UNETR model suggest that further exploration of 3D models, coupled with explainable AI, could yield significant advancements in the field. Future research should focus on enhancing the interpretability of 3D models, which will be crucial for their adoption in real-world clinical applications.

References

[1] A. Abdollahi, B. Pradhan, and A. Alamri, "Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data," *Ieee Access*, vol. 8, pp. 179 424–179 436, 2020.

[2] M. Adewole, J. D. Rudie, A. Gbdamosi, *et al.*, "The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa)," *ArXiv*, 2023.

[3] S. Bakas, H. Akbari, A. Sotiras, *et al.*, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[4] S. Bakas, M. Reyes, A. Jakab, *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.

[5] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-unet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quantitative imaging in medicine and surgery*, vol. 10, no. 6, p. 1275, 2020.

[6] H. Cao, Y. Wang, J. Chen, *et al.*, *Swin-unet: Unet-like pure transformer for medical image segmentation*, 2021. arXiv: 2105.05537 [eess.IV].

[7] M. J. Cardoso, W. Li, R. Brown, *et al.*, *Monai: An open-source framework for deep learning in healthcare*, 2022. arXiv: 2211.02701 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2211.02701.

[8] C. S. Center, *Cancer statistics*. [Online]. Available: https://cancerstatisticscenter.cancer.org/.

[9] S. L. Chau, R. Hu, J. Gonzalez, and D. Sejdinovic, *Rkhs-shap: Shapley values for kernel methods*, 2022. arXiv: 2110.09167 [stat.ML]. [Online]. Available: https://arxiv.org/abs/2110.09167.

[10] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, *3d u-net: Learning dense volumetric segmentation from sparse annotation*, 2016. arXiv: 1606.06650 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1606.06650.

[11] M. D. Cirillo, D. Abramian, and A. Eklund, *What is the best data augmentation for 3d brain tumor segmentation?* 2021. arXiv: 2010.13372 [eess.IV].

[12] I. Csiszar, "*I*-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975. DOI: 10.1214/aop/1176996454. [Online]. Available: https://doi.org/10.1214/aop/1176996454.

[13] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, Apr. 2020, ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2020.01.013. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2020.01.013.

[14] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[15] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, IEEE, 2018, pp. 0210–0215.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2010.11929.

[17] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, *The importance of skip connections in biomedical image segmentation*, 2016. arXiv: 1608.04117 [cs.CV].

[18] D. Duque-Arias, S. Velasco-Forero, J.-E. Deschaud, *et al.*, "On power Jaccard losses for semantic segmentation," in *VISAPP 2021 : 16th International Conference on Computer Vision Theory and Applications*, Vienne (on line), Austria, Feb. 2021. [Online]. Available: https://hal.science/hal-03139997.

[19] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf, *Residual conv-deconv grid network for semantic segmentation*, 2017. arXiv: 1707.07958 [cs.CV].

[20] G. Gerig, M. Jomier, and M. Chakos, "Valmet: A new validation tool for assessing and improving 3d object segmentation," in *Medical Image Computing and*

Computer-Assisted Intervention–MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings 4, Springer, 2001, pp. 516–523.

[21] A. Hatamizadeh, Y. Tang, V. Nath, *et al.*, *Unetr: Transformers for 3d medical image segmentation*, 2021. arXiv: 2103.10504 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2103.10504.

[22] G. Huang, Z. Liu, and K. Weinberger, "Densely connected convolutional networks," p. 12, Aug. 2016.

[23] H. Huang, L. Lin, R. Tong, *et al.*, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2020, pp. 1055–1059.

[24] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*, PMLR, 2019, pp. 285–296.

[25] B. Kim, M. Wattenberg, J. Gilmer, *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 2668–2677. [Online]. Available: https://proceedings.mlr.press/v80/kim18d.html.

[26] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng, *H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes*, 2018. arXiv: 1709.07330 [cs.CV].

[27] Z. Liu, Y. Lin, Y. Cao, *et al.*, *Swin transformer: Hierarchical vision transformer using shifted windows*, 2021. arXiv: 2103.14030 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.14030.

[28] Z. Liu, L. Tong, L. Chen, *et al.*, "Deep learning based brain tumor segmentation: A survey," *Complex & intelligent systems*, vol. 9, no. 1, pp. 1001–1026, 2023.

[29] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, 2015. arXiv: 1411.4038 [cs.CV].

[30] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017. arXiv: 1705.07874. [Online]. Available: http://arxiv.org/abs/1705.07874.

[31] B. H. Menze, A. Jakab, S. Bauer, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[32] Y.-H. Nai, B. W. Teo, N. L. Tan, *et al.*, "Comparison of metrics for the evaluation of medical segmentations using prostate mri dataset," *Computers in biology and medicine*, vol. 134, p. 104 497, 2021.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. arXiv: 1602.04938. [Online]. Available: http://arxiv.org/abs/1602.04938.

[34] O. Ronneberger *et al.*, "Unet: Convolutional networks for biomedical image segmentation," *Journal of Biomedical Imaging*, 2015.

[35] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1505.04597.

[36] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *Machine Learning in Medical Imaging*, Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, Eds., Cham: Springer International Publishing, 2017, pp. 379–387.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019, ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7.

[38] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. arXiv: 1610.02391. [Online]. Available: http://arxiv.org/abs/1610.02391.

[39] M. Sharma and N. Miglani, "Automated brain tumor segmentation in mri images using deep learning: Overview, challenges and future," *Deep learning techniques for biomedical and health informatics*, pp. 347–383, 2020.

[40] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool," *BMC medical imaging*, vol. 15, pp. 1–28, 2015.

[41] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, *Training data-efficient image transformers & distillation through attention*, 2021. arXiv:

2012.12877 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2012.12877.

[42] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102 470, 2022, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102470. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001177.

[43] B. Wang, F. Deng, P. Jiang, S. Wang, X. Han, and H. Zheng, "Witunet: A u-shaped architecture integrating cnn and transformer for improved feature alignment and local information fusion," *arXiv preprint arXiv:2404.09533*, 2024.

[44] Y. Xie, J. Yan, L. Kang, Y. Guo, J. Zhang, and X. Luan, "Fct: Fusing cnn and transformer for scene classification," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 4, pp. 611–618, 2022.

[45] A. Zaki and K. Ali, "Segmentation of human brain gliomas tumour images using u-net architecture with transfer learning," *Diyala Journal of Engineering Sciences*, vol. 15, pp. 17–29, Mar. 2022. DOI: 10.24237/djes.2022.15102.

[46] Y. J. Zhang, "A review of recent evaluation methods for image segmentation," in *Proceedings of the sixth international symposium on signal processing and its applications (Cat. No. 01EX467)*, IEEE, vol. 1, 2001, pp. 148–151.

[47] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[48] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, pp. 3–11.

[49] K. J. Zülch, *Brain tumors: their biology and pathology*. Springer-Verlag, 2013.