3D ATTENTION U-NET BASED MODEL FOR MULTI-REGION BRAIN TUMOR SEGMENTATION

Research Project Phase-2 Report

Submitted by

ANILA KUNJUMON

REG NO : TKM22MEAI04

to

the APJ Abdul Kalam Technological University in partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY
IN
Artificial Intelligence

Under the guidance of
Prof. Chinnu Jacob

Centre for Artificial Intelligence

TKM College of Engineering Kollam

JUNE 2024

Thangal Kunju Musaliar College of Engineering
Centre for Artificial Intelligence

C E R T I F I C A T E

This is to certify that this report titled ***3D ATTENTION U-NET BASED MODEL FOR MULTI-REGION BRAIN TUMOR SEGMENTATION*** is a bonafide record of the **Research project Phase-2** presented by **ANILA KUNJUMON (TKM22MEAI04)**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M. Tech in Artificial Intelligence in APJ Abdul Kalam Technological University.**

Internal Supervisor	**Project Coordinator**	**Head of the Department**
Prof. Chinnu Jacob Assistant Professor Centre for AI TKMCE	Dr. Sumod Sundar Assoc Professor Centre for AI TKMCE	Dr. Imthias Ahamed T P Professor Centre for AI TKMCE

ACKNOWLEDGEMENT

A successful project is a fruitful culmination of efforts by many people, some directly involved and some others indirectly, by providing support and encouragement. Firstly I would like to thank the almighty for giving me the wisdom and grace for making my project a memorable one. I thank him for steering me to the shore of fulfillment under his protective wings.

I express my sincere gratitude to **Dr. T A Shahul Hameed**, Principal of T.K.M College of Engineering for allowing me to present my Research Project Phase-2. I would like to thank **Dr. Imthias Ahamed T P**, Professor and Head of the Department, Centre for Artificial Intelligence, TKM College of Engineering, Kollam, for his constant support and encouragement throughout the work.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my Internal Supervisor **Prof. Chinnu Jacob**, Assistant Professor, Centre for Artificial Intelligence (AI), TKM College of Engineering, Kollam and Project Coordinator, **Dr. Sumod Sundar** Associate Professor, Centre for Artificial Intelligence (AI), TKM College of Engineering, Kollam for their expert guidance, cooperation, and immense encouragement. I also extend my thanks to the entire faculty and staff members of the Centre for AI, TKMCE, who have encouraged me throughout this work.

I also express my thanks to my loving parents and friends for their support and encouragement in the successful completion of this work.

Anila Kunjumon

Abstract

Brain tumors are characterized by the abnormal growth of cells within or near the brain tissues. Accurate segmentation of brain tumors is crucial for effective clinical decision making. The traditional models encounter challenges in accurately delineating tumor regions. In addition, training robust segmentation models on high-resolution magnetic resonance data requires high computational resources. This work proposes 3D attention-based U-Net architecture for multi-region segmentation of brain tumors using a single stacked multi-modal volume. Incorporating the Squeeze and Excitation (SE) attention mechanism into the encoder side of the proposed model facilitates the capture of fine-grained details and the prioritization of significant regions through segmentation areas. This work uses a publicly available Brats2020 (brain tumor segmentation) dataset. The segmentation performance of the model is optimized by evaluation matrices such as Dice coefficient metrics. The proposed model achieved Dice coefficient scores of 0.84, 0.89, and 0.86 for the Enhancing Tumor, the Tumor Core, and the entire Tumor, respectively. This work contributes to accurate brain tumor identification and highlights the transformative role of deep learning in advancing medical image analysis for improved patient care.

Contents

List of Figures

Chapter 1

Introduction

Tumors, which are abnormal cell clusters that form in the human body, can be classified as malignant—cancerous and able to spread to adjacent tissues—or benign, indicating non-cancerous growths. The manual detection of these tumors is a demanding and difficult task for healthcare providers. Consequently, there is an increasing need for smart systems that can autonomously identify cancerous tumors in certain areas of the body. Progress in medical technology has profoundly impacted the diagnosis and prognosis of diseases.

Brain tumors represent a type of abnormal growth within the skull. The complex and sensitive structure of the brain requires the use of non-invasive diagnostic methods, with Magnetic Resonance Imaging (MRI) being the preferred choice. Magnetic resonance imaging generates three-dimensional images of the brain, which can be examined from various angles: coronal, sagittal, and transverse planes, as shown in Figure 1.1. Each view offers crucial insights into the presence of abnormal growths in the skull. The classification of MRI scans according to these planes has been shown to improve the precision of brain tumor detection.

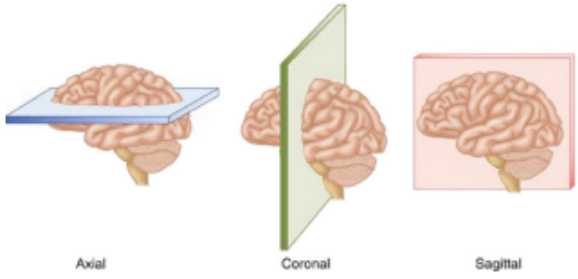

Figure 1.1: One axial slice of an MR image of a high-grade glioma.

Brain tumors that are extremely aggressive, known as gliomas, have high mortality rates and short survival times when they reach advanced stages. The World Health Organization (WHO) classifies tumor malignancy into four grades, from grade I to grade IV, based on histopathology and genomic criteria. Low Grade Gliomas (LGG) are grades I and II, while High Grade Gliomas (HGG) are grades III and IV. LGG tumors progress slowly and often without symptoms, so they are usually diagnosed less frequently than HGG tumors.Although

a biopsy is invasive, it is the gold standard for determining the grade of a tumor. However, this method is prone to sampling errors and discrepancies between observers. Tumors are categorized based on microscopic cell similarities and levels of differentiation. To identify mutations, genotype analysis is combined with histological findings, including the 1p/19q codeletion status and IDH type.

Medical imaging modalities such as X-rays, CT scans, and MRI scans are utilized for disease diagnosis. Among these, Magnetic Resonance Imaging (MRI) is extensively employed for brain tumor detection and treatment. MRI provides high-resolution images of the brain, facilitating both diagnosis and treatment planning for tumors. The segmentation of brain tumors using MRI has significantly improved diagnostic accuracy, treatment efficacy, and growth rate assessment. Accurately delineating tumor regions from MRI images is crucial, given the complex structure of brain tumors, which poses challenges in distinguishing cancerous tissue from healthy brain tissue. Therefore, the development of automated segmentation techniques is essential to achieve more precise and efficient tumor detection and segmentation. Manual segmentation methods are time-consuming and prone to human errors.

These techniques utilize radiomic analysis of gliomas, which involves examining the tumor's shape, heterogeneity, and the lengths of its first and second major axes. Additionally, T1 contrast-enhanced (T1ce) MRI sequences, which highlight features like necrosis and enhancement, can be used for diagnosis.

Gliomas represent the most common primary brain tumors in adults, ranging from mild to severe in their level of aggressiveness and potential outcomes. Magnetic Resonance Imaging (MRI) plays a crucial role in precisely evaluating the diversity of the tumor, commonly employing sequences such as T1 weighted (T1), T1-weighted with gadolinium contrast enhancement (T1Gd), T2 weighted (T2), and fluid-attenuated inversion recovery (FLAIR).

Different regions of tumors, each with distinct features, can be observed on MRI scans. Compared to the T1 sequence, the enhancing tumor (ET) appears more intense in the T1Gd sequence. Conversely, the T1-Gd sequence shows less intense "non-enhancing tumor" (NET) and "necrotic tumor" (NCR) compared to T1. Additionally, in the FLAIR sequence, "peritumoral edema" (ED) appears more pronounced. These subregions are essential for tumor analysis: ET represents the main tumor mass, while the tumor core (TC) comprises ET, NET, and NCR. The addition of ED to TC forms the entire tumor (WT). Figure 1.2, created using the 3D slicer, illustrates each sequence and tumor subvolume.

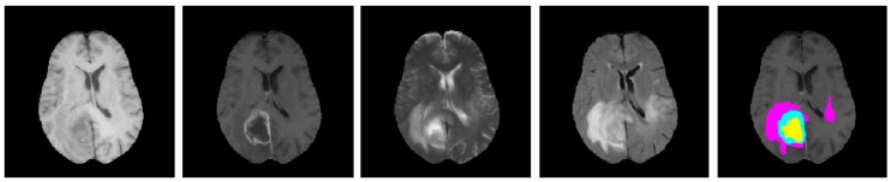

Figure 1.2: Tumor Subvolume

Precisely defining each tumor subregion is crucial for managing the patient's condition, especially after surgery. According to the Radiation Therapy Oncology Group (RTOG) guidelines, radiation oncologists need to identify the tumor and describe details about the surgical resection cavity, remaining enhancement tumor, and surrounding edema. Accurate segmentation meets clinical needs and allows for prognostic factor discovery through advanced methods like deep learning or radiomics. These modern techniques provide valuable insights into prognosis and help tailor treatment plans to improve patient outcomes.

The 2020 Multimodal Brain Tumor Segmentation Challenge consisted of three main tasks. Firstly, participants were required to segment various tumor subregions. Secondly, they had to predict overall patient survival (OS) using preoperative magnetic resonance imaging. Lastly, participants needed to assess segmentation uncertainty measures.The main goal of the Segmentation challenge was to accurately identify the enhancing tumor (ET), tumor core (TC), and whole tumor (WT) regions. The Dice Similarity Coefficient (DSC) was the primary metric used to assess the overlap between two sets, which is a commonly used metric in this context.

The primary challenge encountered in both domestic and international contexts during the segmentation of brain tumors lies in ensuring the accuracy of the segmented images. Traditional segmentation methods, such as threshold segmentation and region segmentation, rely on algorithms similar to those described in the conventional segmentation method. However, these methods, including the Fuzzy C-Means (FCM) algorithm, suffer from drawbacks like time-consuming processes and defects. These issues pose challenges to ensuring efficiency and quality in processing brain tumor images. Addressing the limitations of existing algorithms in brain tumor image detection, a U-Net-based segmentation algorithm has been developed. This U-Net-based algorithm overcomes drawbacks by achieving efficient training, robustness, and accurate and complete segmentation of brain tumor images. Using image recognition and deep learning techniques, the U-Net network's algorithm assists physicians in disease localization, condition analysis, and diagnosis assistance without the need for manual intervention. This not only improves productivity but also holds significant clinical implications for guiding diagnosis and treatment processes.

1.1 Radiomic analysis of anatomical MR images

Perfusion imaging and MR spectroscopy provide crucial information for glioma grading, but anatomical MR imaging is also valuable in this regard. Features like necrosis, mass effect, and cyst formation are known markers of high malignancy. The use of Inversion Recovery

sequences such as FLuid Attenuated Inversion Recovery (FLAIR) enhances lesion and edema detection by canceling the signal of specific tissues like cerebrospinal fluid. Neoangiogenesis, visible with contrast-enhanced sequences like T1ce, is a marker for high-grade gliomas, though not all high-grade gliomas exhibit enhancement signals. Some studies have linked genotype and MR phenotype on anatomical imaging according to WHO 2016 guidelines. For instance, characteristics such as the T2-FLAIR mismatch sign or the sharpness of tumor borders have been associated with 1p/19q codeletion or IDH mutation. Therefore, anatomical MRI can provide information similar to the WHO classification of gliomas.

Chapter 2

Literature Survey

In 1978, EMI Laboratories made a groundbreaking achievement by pioneering the assessment of brain MRI images. This milestone marked the beginning of quantitative analysis in neuroimaging. Since that time, the field has experienced rapid advancements, particularly in automating the detection of various brain disorders. These conditions comprise degenerative diseases, cancer, Alzheimer's disease, schizophrenia, and multiple sclerosis (MS).

A significant aspect of this progress has been the focus on identifying regions of interest (ROIs) within brain images. Identifying ROIs is crucial for a number of tasks in neuroimaging. One important task is evaluating tissue atrophy, which involves measuring the loss of brain tissue over time. Another key task is segmenting different brain tissues, which helps in distinguishing between various types of tissues within the brain. Additionally, measuring correlations among these tissues is vital for understanding how different regions of the brain interact and are affected by diseases.

Overall, these advancements in identifying and analyzing ROIs within brain images have greatly enhanced our ability to diagnose and understand brain disorders. This progress has paved the way for more precise and automated methods in neuroimaging, contributing to improved patient outcomes.

Traditionally, brain tumor segmentation relied on techniques such as support vector machines (SVMs) and random forests. These methods required the use of manually crafted features, which were designed by experts to help the algorithms identify tumors.In 2015Ronneberger introduced a new deep learning-based method called U-Net in 2015. U-Net stands

out because it is uniquely designed with pathways and skip connections. These design elements enhance the network's performance by allowing it to learn more effectively from the data. Unlike traditional methods, U-Net does not require manual feature design, which makes it more flexible and powerful. Moreover, U-Net incorporates mirroring to precisely

predict border pixels. This means that the model can better identify the edges of tumors, which is crucial for accurate segmentation. The adoption of U-Net has significantly improved the accuracy and efficiency of automated brain tumor segmentation. It allows for more precise and reliable detection of brain tumors, which is essential for effective diagnosis and treatment planning. This has been a major advancement in the field of medical image

analysis, leading to better outcomes for patients.

The success of AlexNet in 2012, there has been a surge in efforts to improve convolutional networks' performance. Some methods focus on altering the connectivity patterns of convolutional, pooling, and fully connected layers, while others aim to increase the network's depth. These strategies have enhanced the breadth and depth of network performance.

Zhou et al.[1] introduced UNet++, which enhances UNet by stacking multiple smaller UNet units and utilizing a pruning algorithm to remove unnecessary connections during testing. Unlike 2D UNet, 3D UNet can directly process full 3D images without requiring them to be sliced beforehand.

Dong et al.[2] introduced a fully convolutional network based on the U-Net architecture for brain tumor detection and segmentation. However, accurately identifying the enhancing region in LGG cohorts posed challenges for this network. Havaei et al.[3] enhanced brain tumor segmentation performance on the BraTS dataset by utilizing two-pathway and cascaded architectures in their network. The two-pathway architecture captures both local and global features to predict pixel labels, considering both the patch's location in the brain and visual details. Wang et al.[4] expanded on the cascaded architecture by incorporating cascaded deep CNN for efficient brain tumor segmentation and anisotropic CNN for automatic brain tumor segmentation.

Noh et al.[5] introduced an encoder-decoder network for semantic segmentation in natural scenes, utilizing multiple convolution and pooling layers to convert high-resolution features into low-level prominent edges. Addressing the limitations of Simonyan et al.[6] proposed SegNet, a revised encoder-decoder architecture that efficiently stores data using max-pooling indices, eliminating the need for precise feature map storage.

Kamnitsas et al. [7] introduced a 3D CNN with a fully connected conditional random field (CRF) for precise brain lesion segmentation. This 11-layer deep architecture processed adjacent image patches in a single pass, adapting to inherent class imbalances and incorporating a dual pathway architecture for multi-scale processing. Cui et al. [8] proposed a deep cascaded neural network comprising a tumor localization network (TLN) and an intra-tumor classification network (ITCN) for brain tumor segmentation. Lin et al. [9] utilized dense CRF learning in conjunction with CNNs for segmentation refinement. Zhao et al. [10] enhanced tumor segmentation by integrating a fully convolutional neural network (FCNN) with CRF. These techniques employ a patch-based approach, dividing medical scans into patches for testing and training. Consequently, there is a clear need for a model that combines both local and global features and accepts full images as input to tackle the problem of class imbalance.

The attention mechanism is inspired by the human visual system, enabling individuals to focus on specific areas of interest while disregarding others, aiding in the suppression of irrelevant details and obtaining detailed information about the object of interest. Combining Convolutional Neural Networks (CNNs) with attention mechanisms has shown promising results across various fields. For instance, Trebing et al.[11] introduced SmaAt-UNet, a weather forecast prediction system that merges the Convolution Block Attention Module

(CBAM) with UNet's attention mechanism. Similarly, Li et al.[4] achieved favorable outcomes by incorporating lightweight attention mechanisms with UNet for segmenting retinal vessel images and establishing connections between feature images using the attention based mechanism. The neural network learns crucial information from each image. Subsequently, the network automatically adjusts weights to suit different recognition tasks. The attention mechanism's main role is to identify important information and suppress irrelevant details, aiding in the rapid extraction of useful information from vast datasets and enhancing feature extraction capabilities.

To develop a superior 3D U-Net architecture tailored for precise tumor segmentation in masked MRI brain images, we can integrate insights from the aforementioned studies on attention mechanisms and neural network architectures. Firstly, we can incorporate elements from the CBAM (Channel Attention Module) which effectively combines both channel and spatial attention mechanisms. This allows the network to capture both inter-channel relationships and spatial dependencies, enhancing feature extraction capabilities.Additionally, we can leverage the lightweight and parameter-efficient ECA (Efficient Channel Attention) module to further refine the attention mechanism in our 3D U-Net architecture. By dynamically adjusting the convolutional kernel size based on the number of channels, the ECA module can adaptively capture channel-wise dependencies, leading to improved segmentation accuracy.Furthermore, we can fine-tune the training techniques by employing advanced optimization algorithms such as Adam optimizer and exploring techniques like learning rate scheduling and data augmentation. These techniques can help improve the convergence speed and generalization ability of the model.

Chapter 3

Methodology

3.1 Objective(s)

- To design 3D attention-based U-Net architecture to segment brain tumors across multiple regions using a single stacked multi-modal volume.
- To utilize Squeeze-and-Excitation (SE) attention mechanism improvements the feature channels and improves the accuracy of the network.
- To compare the proposed work with existing works.

3.2 Proposed Methodology

Figure 3.1 depicts the proposed methodology.

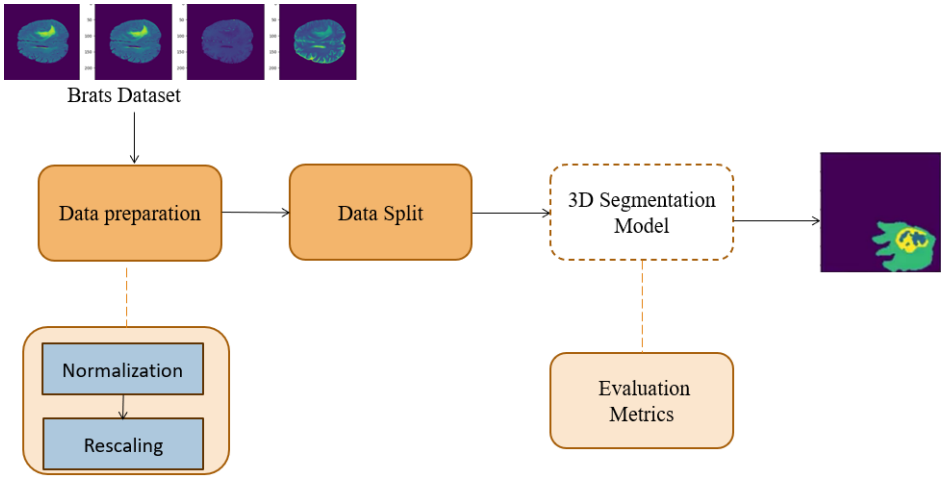

Figure 3.1: Proposed Methodology

The dataset used for brain tumor segmentation is BraTS 2020, acquired from Kaggle. Before training, the images underwent rescaling and normalization. The dataset contains detailed information on 369 patients, with 80% allocated for training, 20% for validation, and the seperate data for testing. Segmentation is performed using a 3D U-Net model, and

model evaluation employs metrics such as Intersection over Union (IoU) and Dice coefficient. Training employs the ADAM optimizer with a learning rate of 0.001. The model's ability to generalize, its accuracy, and overall performance are assessed and tested on the validation set.

3.3 Multimodal Brain Tumor Segmentation

The Multimodal Brain Tumor Segmentation Challenge 2020 (BraTS 2020) was divided into three distinct tasks aimed at advancing the field of brain tumor analysis using MRI scans. These tasks were:

- **Segmentation of Different Tumor Sub-regions:** The primary objective of this task was to precisely identify the enhancing tumor (ET), tumor core (TC), and whole tumor (WT) regions within the brain tumor. Accurate segmentation is crucial for treatment planning, clinical decision-making, and monitoring disease progression. Participants were required to develop algorithms capable of accurately identifying and segmenting these critical sub-regions, each exhibiting distinct biological behavior and responses to therapy.

- **Predicting Patient Overall Survival (OS):** This task involved predicting the overall survival of patients based on pre-operative MRI scans. This predictive task posed a significant challenge as it required the development of models that could interpret complex imaging data to forecast patient outcomes. The ability to predict overall survival from imaging data alone has profound implications for personalizing treatment strategies, optimizing therapeutic approaches, and improving patient counseling.

- **Assessing Uncertainty Measures in Segmentation:** This task aimed at evaluating and quantifying the uncertainty in segmentation outputs. Understanding and effectively communicating the confidence in segmentation results are crucial for their application in clinical settings. Reliable uncertainty measures can help clinicians gauge the trustworthiness of automated segmentations, potentially guiding further diagnostic procedures and treatment decisions.

In the segmentation challenge, participants were tasked with accurately identifying the enhancing tumor (ET), the tumor core (TC), and the whole tumor (WT) sections of tumor. To comprehensively evaluate segmentation performance, both overlap measures and distance metrics were employed as the primary evaluation metrics. This approach ensured a thorough assessment of segmentation performance from multiple perspectives, facilitating a more robust evaluation of algorithm effectiveness.

A key metric used for evaluating segmentation accuracy was the Dice Similarity Coefficient (DSC). The DSC is a widely used measure that quantifies the overlap between two

sets. In the context of comparing predicted segmentation with ground truth, the DSC can be defined mathematically as follows:

$$\text{DSC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \tag{3.1}$$

where TP represents the true positives (the number of correctly classified voxels), FP denotes the false positives, and FN signifies the false negatives. The numerator $2 \times \text{TP}$ accounts for the number of correctly identified voxels in the intersection of the predicted and ground truth sets. The denominator $2 \times \text{TP} + \text{FP} + \text{FN}$ represents the total number of voxels in both the predicted and ground truth sets. This normalization ensures that the DSC ranges between 0 and 1, where a value of 1 indicates perfect overlap and 0 indicates no overlap at all. This metric is particularly beneficial for assessing the accuracy of segmentations within a specific region of interest, such as a tumor, as it remains unaffected by the extent of background present in the image.

In addition to the Dice Similarity Coefficient (DSC), the Hausdorff distance served as another crucial assessment metric. It measures the maximum distance between the margins of two contours, providing a stringent evaluation of segmentation accuracy. The Hausdorff distance is particularly sensitive to outliers: even a single voxel deviation from the reference segmentation can result in a significantly high Hausdorff distance, regardless of overall voxel overlap. This metric is invaluable when assessing the clinical significance of a segmentation. For instance, while the Dice metric may indicate satisfactory overall overlap, the Hausdorff distance can highlight instances where manual correction by a radiation oncologist is necessary, such as when distant healthy brain tissue is erroneously included in the tumor segmentation, aiming to prevent adverse outcomes for the patient.

By combining the DSC and Hausdorff distance, the challenge provided a robust framework for evaluating how well the algorithms performed in segmenting the various tumor sub-regions. The DSC focused on overall overlap accuracy, while the Hausdorff distance highlighted the presence of significant outliers. Together, these metrics ensured that automated segmentation tools could be robustly assessed and improved, ultimately contributing to more reliable and clinically useful outcomes in brain tumor analysis. This comprehensive evaluation method helped push the boundaries of automated brain tumor segmentation, fostering advancements in the precision and utility of computer-aided diagnostic tools, which are essential for enhancing patient care and treatment outcomes.

3.4 Techniques Used

The following are the techniques used in this work.

3.4.1 Dataset

Three cohorts make up the BraTS 2020 challenge dataset: Testing, Validation, and Training. Every one of these cohorts is essential to the assessment and advancement of segmentation algorithms for the analysis of brain tumors.

The Training dataset includes multi-parametric MRI (mpMRI) scans from 369 patients with diffuse glioma, featuring four types of sequences in each set. native T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid-Attenuated-Inversion-Recovery (FLAIR). These sequences provide comprehensive imaging data that capture different aspects of the tumor's structure and pathology. To ensure consistency and comparability, all MRI volumes undergo several preprocessing steps. These include skull-stripping to remove non-brain tissues, alignment with a standard anatomical atlas to standardize spatial orientation, and resampling to a 1mm voxel resolution for uniformity. The training dataset contains the ground truth (GT) labels are annotated by expert human annotators, comprising three classes: enhancing tumor (ET), tumor core (TC), and whole tumor (WT).

The BraTS 2020 Validation data includes 125 cases of diffuse glioma patients. Similar to the training set, it includes four distinct mpMRI sequences for each case: T1, T1ce, T2, and FLAIR. The validation cohort provides an opportunity for participants to assess their models on previously unseen data, which is crucial for evaluating the generalization performance of the developed algorithms. However, the ground truth labels for the validation data are not disclosed to the participants. This setup simulates a real-world scenario where the model must make accurate predictions without prior knowledge of the true labels.

The BraTS 2020 Testing cohort is reserved for the final ranking of participating teams and comprises 166 cases. This cohort is used to determine the ultimate performance of the algorithms developed by the teams. The specific type of glioma in the testing cases remains undisclosed to the participants, adding an additional layer of challenge and realism. Each team is given a 48-hour window to submit their results for conclusive evaluation on the challenge platform. This time constraint ensures that the algorithms are efficient and capable of processing the data within a practical timeframe.

The BraTS 2020 challenge dataset is meticulously structured to test the robustness, accuracy, and generalizability of brain tumor segmentation algorithms. The training dataset allows for model development and fine-tuning, the validation dataset provides a means to assess performance on new data, and the testing dataset serves as the final arbiter of algorithm effectiveness, ensuring a comprehensive evaluation framework for the participating teams.

3.4.2 Data Preprocessing

The BraTS-2020 dataset, comprising preoperative MRI scans from multiple institutions, is used to test and evaluate the proposed model. Figure 3.2 shows a sample of the data from this dataset. BraTS 2020 aims to segment the different forms, appearances, and histologies of

brain tumors, particularly gliomas. Additionally, the dataset seeks to assess the clinical relevance of segmentation by focusing on predicting overall patient survival through integrative analysis using radiomic features and machine learning algorithms.

The dataset comprises four modalities of 3D brain MRIs from MRI exams, each with dimensions of 128 × 128 × 128 voxels. These modalities include native (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) images. Additionally, the dataset provides a ground truth segmentation map that classifies each voxel into one of four semantic labels: necrotizing and non-enhancing tumor core (NCR/NET), peritumoral edema (ED), unlabeled volume, and GD-enhancing tumor (ET). Each imaging dataset has been meticulously segmented and validated by expert neuroradiologists.

The BraTS 2020 dataset is meticulously structured to provide comprehensive data for developing and testing brain tumor segmentation algorithms. By including multiple MRI modalities, the dataset captures various aspects of tumor biology and pathology, essential for creating robust segmentation models. The integration of radiomic features and machine learning algorithms aims to enhance the predictive accuracy for patient outcomes, making the dataset highly relevant for clinical applications.

The segmentation task within the BraTS 2020 challenge involves accurately identifying and classifying different tumor sub-regions. The provided ground truth segmentation maps, created by experienced neuro-radiologists, ensure high-quality data for training and evaluating algorithms. The four categories in the segmentation maps—Unlabeled volume, Necrotic and non-enhancing tumor core (NCR/NET), Peritumoral edema (ED), and GD-enhancing tumor (ET)—cover the critical regions of interest in glioma analysis.

Moreover, the BraTS 2020 dataset's focus on predicting overall patient survival through radiomic features and machine learning algorithms adds a significant layer of clinical relevance. This integrative approach aims to leverage the detailed imaging data to make more accurate and meaningful predictions about patient outcomes, potentially improving treatment planning and personalized medicine approaches. The BraTS dataset, particularly the 2020 edition, is a comprehensive resource for advancing brain tumor segmentation and prediction models. It provides high-quality, multi-institutional MRI data, expertly annotated segmentation maps, and a framework for integrating radiomic features with machine learning to enhance the clinical applicability of the developed algorithms.

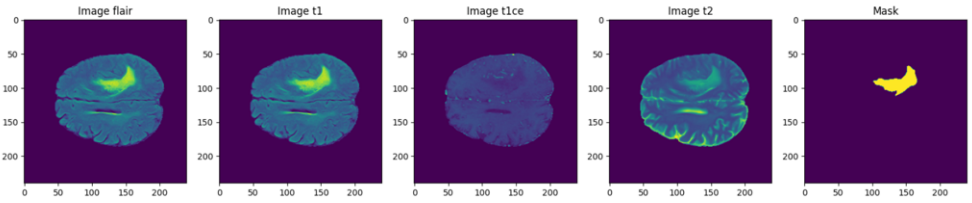

Figure 3.2: Sample data from BraTS dataset

To standardize the input images and mitigate variations in intensity introduced by factors such as device manufacturer, acquisition parameters, and sequence, the mean pixel value is subtracted and divided by the pixel standard deviation values. This normalization ensures consistency in intensity values across scans and patients, thereby facilitating more effective learning and comparison of data by the algorithms.

The main focus is on processing weighted images such as T1ce, FLAIR, and T2 to extract more features during training. T1-weighted and T2-weighted images provide valuable information that aids in prognostic evaluation. Specifically, T1ce-weighted post-contrast images reveal differences in T1 relaxation times between tissues and highlight hypervascular pathologies. These images are particularly useful for identifying regions with abnormal blood-brain barrier permeability, which often indicates tumor activity or other pathological processes.

T2-weighted images are particularly useful for diagnosing inflammatory conditions affecting the brain and spinal cord, as they show hyperintense signals in areas of inflammation, especially near cerebrospinal fluid spaces. These images are crucial for identifying edema and other fluid-related abnormalities by providing essential information about tissue water content. Prioritizing the preprocessing of T1ce, T2, and FLAIR-weighted images enhances feature extraction during training, improving the accuracy of glioma segmentation and prognosis.

Multi-class segmentation Description

The imaging datasets used in this study have undergone manual segmentation by one to four raters, who followed a consistent annotation protocol. These annotations were then reviewed and approved by experienced neuro-radiologists. The segmentation includes three main regions: the GD-enhancing tumor (ET - label 4), peritumoral edema (ED - label 2), and the necrotic and non-enhancing tumor core (NCR/NET - label 1) as shown in Figure 3.1. This annotation protocol aligns with the guidelines outlined in both the BraTS 2012-2013 TMI paper and the latest BraTS summarizing paper. The provided data have undergone preprocessing steps, including co-registration to a standardized anatomical template, interpolation to a uniform resolution and skull-stripping.

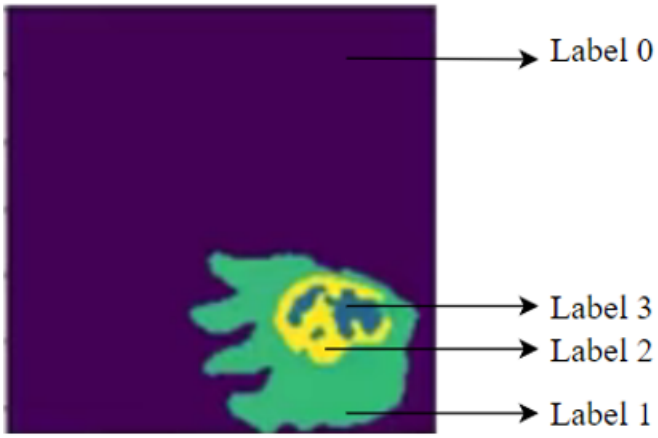

Figure 3.3: Segmentation classes

3.4.3 3D U-Net

The U-Net architecture comprises two main components: downsampling and upsampling layers. Its fundamental structure resembles that of a typical convolutional neural network, with two sets of repeated 3 × 3 convolutions, each followed by a rectified linear unit (ReLU) activation function, and a 2 × 2 pooling operation with a stride of 2 for downsampling. This process effectively doubles the number of image channels after each step in the first section. In the second section, there is a 2 × 2 convolutional upsampling layer, followed by two 3 × 3 convolutions, each with a ReLU activation, and concatenation with the corresponding feature map from the first segment.

Despite having limited data, U-Net has showcased remarkable performance across various real-world applications, particularly in biomedical image processing. Its capability to capture object scales and its consistent downsampling make it well-suited for the complexities inherent in numerous image processing tasks. Notably, the incorporation of "skip-connections" significantly enhances its performance compared to traditional autoencoders. We have opted for U-Net as the foundational architecture for our proposed 3D U-Net due to its established effectiveness in tasks like brain tumor segmentation. This three-dimensional adaptation retains the core U-Net architecture, featuring encoder and decoder paths, as depicted in Figure 3.4.

Upon processing input images sized at 128 × 128 × 128 pixels, the model generates outputs of identical dimensions. Accelerating the training process involves pre-training the encoder to recognize features like edges, textures, and lines, thereby enhancing its functionality. Training extends over 60 epochs with a batch size of 2 and a 4-channel input sized at 128 × 128 × 128 voxels. To optimize training effectiveness, a customized loss function,

encompassing total weighted loss within the model, is employed.

During training, images are cropped to 128 × 128 × 128 voxels, and a constant batch size of 2 is retained across 30 epochs. Optimization is carried out using the Adam optimizer with a learning rate of 0.0001. Additionally, employing the sigmoid activation function in the final convolutional layer proves beneficial for obtaining segmentation results in binary image models, particularly for brain tumor segmentation.

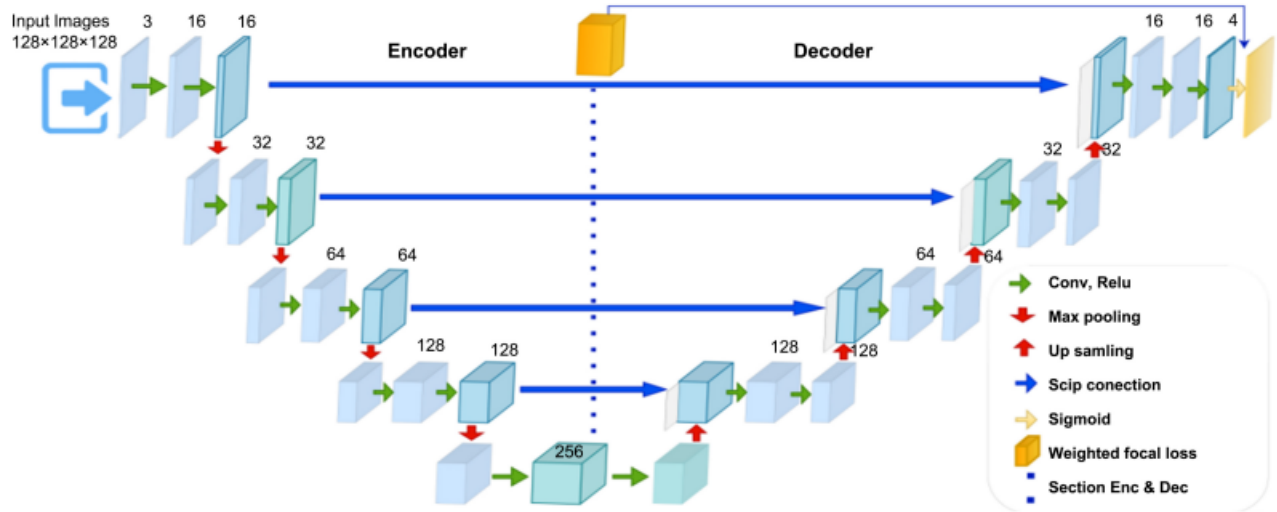

Figure 3.4: Architecture of 3D U-Net

3.4.4 Attention mechanism

Attention is vital for human perception. The human visual system does not process an entire scene all at once; instead, it focuses on it in parts. By taking a series of partial glimpses and concentrating on key areas, the visual structure is captured more effectively. This selective focus allows humans to navigate complex environments efficiently, distinguishing important details from irrelevant background information.

Building on the understanding of how attention enhances human perception, researchers have recently made various efforts to incorporate attention processing into Convolutional Neural Networks (CNNs) to improve their performance in large-scale classification tasks. One notable example is the Residual Attention Network, which employs an encoder-decoder style attention module. This innovative approach allows the network to refine feature maps effectively, resulting in improved performance and robustness to noisy inputs. By refining feature maps, the network can better distinguish between relevant and irrelevant information, much like the human visual system.

Instead of directly computing the 3D attention map, the Residual Attention Network

learns channel attention and spatial attention separately. This division of attention processing is advantageous because it reduces computational and parameter overhead. The separate generation of attention for the 3D feature map is less resource-intensive, making the module more efficient. Consequently, this efficient attention mechanism can be used as a plug-and-play module, easily integrated into pre-existing base CNN architectures. This modularity is beneficial for researchers and practitioners, as it allows for the enhancement of existing networks without significant redesigns. The integration of attention mechanisms into CNNs represents a significant advancement in neural network design, drawing inspiration from the human visual system to enhance machine perception and performance.

The Squeeze-and-Excitation (SE) module, a pioneering technique in deep learning, leverages global average-pooled features to compute channel-wise attention. However, recent studies have revealed that relying solely on these average-pooled features might not yield optimal results, particularly when it comes to fine-tuning channel attention. To address this limitation, researchers have advocated for the incorporation of max-pooled features alongside the average-pooled ones, enriching the SE module's ability to discern and prioritize relevant channels effectively.

Moreover, while the SE module excels in channel-wise attention, it overlooks spatial attention, a critical aspect in determining the precise "where" to focus within an image. Recognizing this gap, the Convolutional Block Attention Module (CBAM) emerges as an innovative solution. CBAM adopts a comprehensive approach by integrating both spatial and channel-wise attention mechanisms into its architecture. This holistic strategy enables the model not only to identify important channels but also to pinpoint their spatial locations, enhancing its overall understanding of visual features.

Empirical evaluations have unequivocally demonstrated the superiority of CBAM over its predecessors. By harnessing both spatial and channel-wise attention, CBAM consistently outperforms models relying solely on channel-wise attention in various tasks, including image classification and object detection. This empirical evidence underscores the efficacy and practical relevance of integrating multi-modal attention mechanisms into deep learning architectures, paving the way for more robust and versatile artificial intelligence systems.

The attention mechanism played a crucial role in feature extraction and refinement within the depicted Figure 3.5. Comprising two stages - channel attention and spatial attention - this mechanism proved instrumental in achieving impressive segmentation results, even surpassing those obtained with a 3D UNet model, despite utilizing a simpler 2D UNet baseline. By effectively suppressing noise and irrelevant features while simultaneously enhancing feature map resolution, this approach demonstrated its efficacy. Moreover, it boasted a parsimonious parameter count compared to alternative models.

Researchers explored various integration strategies for attention gates within the UNet architecture. Some opted for embedding these gates into skip connections, while others incorporated them directly into the decoder stage. Regardless of the specific approach, the integration of attention mechanisms facilitated the production of refined segmentation maps across different resolutions. These refined maps were then aggregated in the final stage of

the segmentation process, resulting in superior segmentation performance.

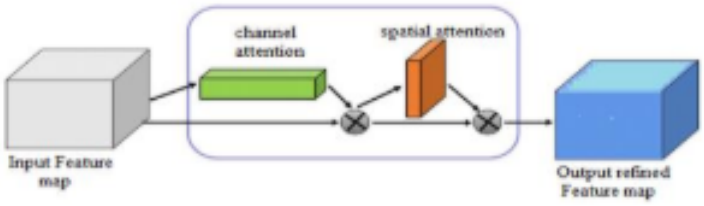

Figure 3.5: Attention module

The channel attention stage operates by refining feature maps individually on a per-channel basis, while the spatial attention stage enhances feature responses within specific regions of interest, such as the brain region in tasks like brain segmentation. Recent studies have seen attention modules integrated into cascaded architectures to refine feature maps after each stage, leading to improved segmentation outputs. This iterative refinement of feature maps at the end of each stage has proven instrumental in achieving enhanced segmentation results. Furthermore, researchers have explored integrating attention modules into a variety

of architectures beyond traditional segmentation models. These include autoencoders and Generative Adversarial Networks (GANs), aiming to boost their performance by leveraging attention mechanisms. Such integrations showcase the versatility and effectiveness of attention mechanisms across different types of neural network architectures.

Channel Attention Module

To construct a channel attention map, this approach involves examining the intricate relationships among different channels of features. Essentially, each channel within a feature map functions akin to a specialized detector, honing in on specific features embedded within an input image. In this endeavor to efficiently compute channel attention, embark on compressing the spatial dimensions of the input feature map. While the conventional practice often resorts to average-pooling for the aggregation of spatial information, recent advancements in research have shed light on the value that max-pooling brings to the table. It has been increasingly recognized that max-pooling, too, holds the potential to glean invaluable insights into the distinct features of objects. Hence, the dual-pronged approach integrates both average-pooled and max-pooled features. The meticulous experiments have unequivocally demonstrated that the fusion of these features serves to markedly enhance the representation prowess of networks. This notable enhancement stands in stark contrast to the isolated utilization of each feature, thereby underscoring the resounding efficacy of this methodology. Figure 3.6 depicts the channel attention Module.

First, we aggregate spatial information of a feature map by employing both average-pooling and max-pooling operations. This generates two different spatial context descriptors:

F^c_{avg} and F^c_{max} representing average-pooled features and max-pooled features, respectively. Subsequently, both descriptors are fed into a shared network to produce our channel attention map $M_c \in R^{C\times1\times1}$ This shared network comprises a multi-layer perceptron (MLP) with one hidden layer. To reduce parameter overhead, the hidden activation size is set to $r \times 1 \times 1$.

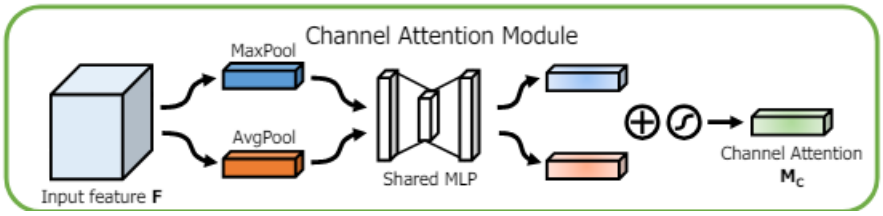

Figure 3.6: Channel Attention Module

Spatial Attention Module

A spatial attention map by utilizing the inter-spatial relationships within the feature maps. This approach is different from channel attention, which focuses on identifying 'what' features are important. Instead, spatial attention is concerned with 'where' the important parts are located in the feature map. This focus on location makes spatial attention complementary to channel attention, as it adds another layer of detail to the feature selection process. Figure 3.7 depicts the spatial attention module.

To compute the spatial attention, we begin by performing two types of pooling operations along the channel axis: average-pooling and max-pooling. These pooling operations are effective in emphasizing the most informative regions of the feature map by reducing dimensionality while retaining crucial spatial information. The results of these pooling operations are two separate 2D maps. The first map, $F^s_{avg} \in R^{1\times H\times W}$, contains the average-pooled features, representing the average value of each spatial location across all channels. The second map, $F^s_{max} \in R^{1\times H\times W}$, consists of the max-pooled features, capturing the maximum value at each spatial location across all channels.

Once these two 2D maps are generated, we concatenate them along the channel dimension. This concatenation effectively combines the information from both the average-pooled and max-pooled features, creating a comprehensive feature descriptor that highlights key spatial regions. The combined descriptor is then fed into a convolution layer, which processes this information to produce the final spatial attention map $M_s(F) \in R^{H\times W}$.

The convolution layer applied to the concatenated feature descriptor is a standard convolution operation. It is responsible for learning the weights that will determine the areas of the feature map to emphasize or suppress. The output of this convolution operation, the spatial attention map, encodes information about which spatial locations in the feature map

are important for the task at hand.

This spatial attention mechanism effectively guides the model to focus on the most relevant parts of the input, enhancing the model's ability to interpret and utilize spatial information. By combining the strengths of both spatial and channel attention, the model can achieve a more nuanced understanding of the input data, leading to improved performance on various tasks.

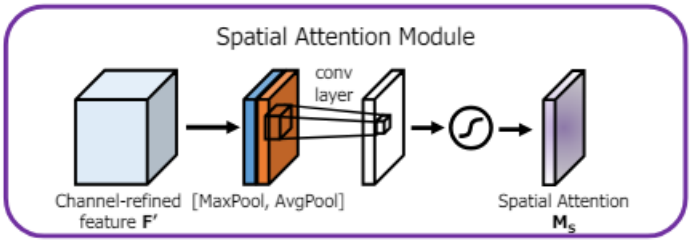

Figure 3.7: Spatial Attention Module

3.4.5 Proposed Model

The proposed network architecture enhances the original 3D-UNet architecture depicted in Figure 3.4 by integrating SE (Squeeze-and-Excitation) modules. Due to GPU resource constraints, the number of layers in both the upsampling and downsampling stages had to be reduced from four to three.

Downsampling layers play a crucial role in extracting contextual information and enhancing the classifier's classification performance by capturing intricate features. Conversely, upsampling layers aim to enlarge the original input. The 3D-UNet architecture follows a U-shaped design, comprising encoding and decoding sections. Within the encoding section, each operation employs two convolution blocks with a 3 x 3 x 3 kernel size. Subsequently, the feature map is passed to the SE module located on the left. Following this, the pooling operation, which employs max-pooling to compare parameters and extract crucial features, is linked to the output of the attention module. The size of the pooling step is set to two by two by two. A corresponding decoding section follows the encoding section.

The decoding section mirrors the encoding section with three upsampling layers. A pivotal aspect of the UNet network is the skip connection mechanism, which amalgamates low-level and high-level semantic information to capture more meaningful features. At each layer in the encoding section, the corresponding upsampling output and attention output are concatenated. Subsequently, the feature channel undergoes a division into two using two convolution operations, maintaining a kernel size of 3 x 3 x 3. The final layer's output is processed through a sigmoid function and a 1 x 1 x 1 convolution operation, ensuring that

every pixel in the resulting image falls within the range of 0 to 1. Ultimately, the network model maintains equal input and output sizes.

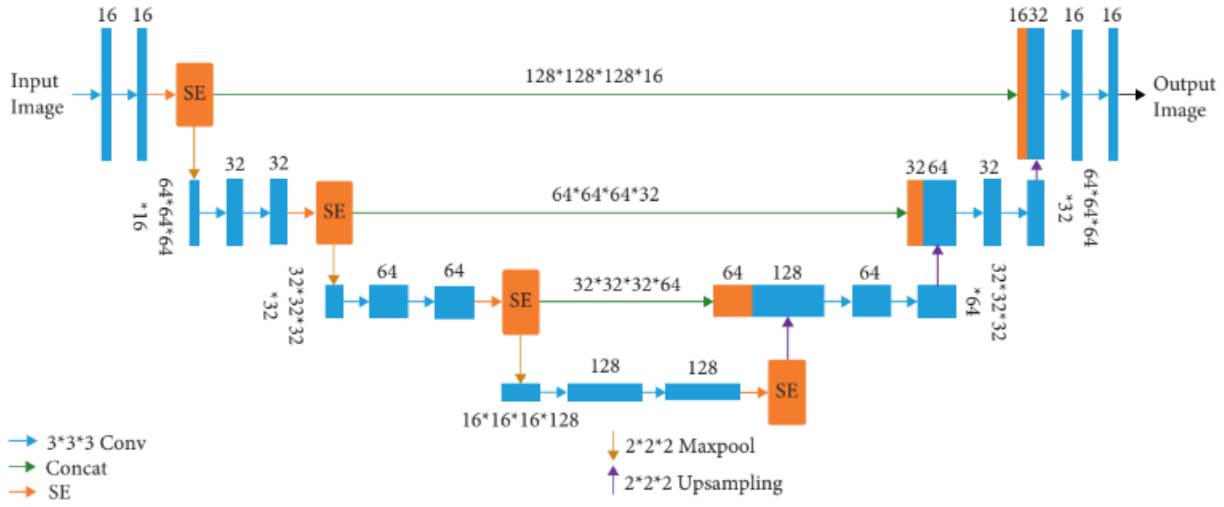

Figure 3.8: Proposed model Architecture

3.4.6 Squeeze-and-Excitation (SE) Module

The SENet module operates by assessing the significance of individual channels within feature maps. Initially, global average pooling compresses the spatial dimensions of each output channel, resulting in a singular scalar value denoting its importance. This approach condenses global feature information into a single value, effectively encompassing a broad receptive field. To maintain consistency between the output dimension and the number of input feature channels, the dimensions transition from H x W x C to 1 x 1 x C.

Following this, the global features undergo processing using a structure resembling FC-ReLU-FC-Sigmoid, resulting in C scalar values ranging between 0 and 1. Initially, the first fully connected (FC) layer reduces dimensionality by a factor of 1/r relative to the input, where r represents the reduction ratio parameter. Subsequently, a second FC layer is employed to restore the original feature dimension. The integration of ReLU activation functions with FC layers contributes to parameter reduction and enhances nonlinearity. Finally, each channel's weight is represented by C values between 0 and 1, obtained through the application of the sigmoid function.

The original feature map undergoes modulation to generate a new feature map using the channel-wise weights obtained from the squeeze-excitation operation. Initially, the squeeze operation employs global average pooling to reduce the feature map dimensions to a single

scalar for each channel. Subsequently, two fully connected (FC) layers and ReLU activation functions are applied to produce weight values between 0 and 1. The resulting squeezed features undergo nonlinear transformation through a fully connected neural network operation called excitation applied to the output of the squeeze operation. As a result, each channel feature is assigned a score by the squeeze-excitation operation, enabling the network to emphasize channels containing significant information while downplaying those with comparatively less importance.

Modifying the SE module to accommodate three-dimensional seismic data requires adjustments from its original design for two-dimensional data. Figure 3.5 illustrates the structure of the SE module. Initially, the squeeze operation compresses the dimensions of H x W x Z x C into a scalar of 1 x 1 x 1 x C using global average pooling. The formula for this operation is expressed as:

$$z_c = \frac{1}{XYZ} \sum_{i=1}^{X} \sum_{j=1}^{Y} \sum_{t=1}^{Z} u_c(i,j,t), \quad z \in R^c \tag{3.2}$$

In this formula, the original feature is represented by u_c, while the dimensions of the seismic input are denoted by X, Y, and Z.

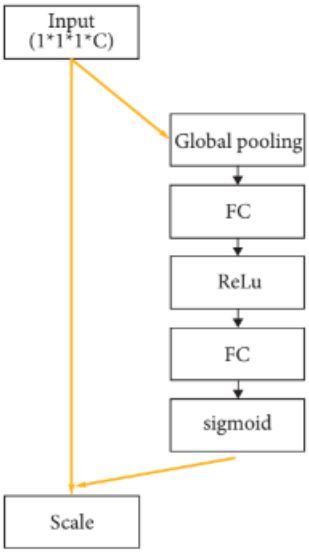

Figure 3.9: A flow chart for the SE Module.

Next, the excitation operation, as indicated by equation (3.2), involves the first FC layer, which is represented by W1 multiplied by z. The dimension of W1 is C/r x C, where r is the scaling parameter set to 8 in this study to reduce parameters. Multiplying W1 by z results in an output dimension of 1 x 1 x 1 x C/r, which remains the same. The second FC layer is then obtained by multiplying W2 by the output of the activation function, with W2 having

a dimension of C x C/r. Consequently, the output dimension is 1 x 1 x 1 x C. Finally, C scalars between 0 and 1 are acquired through the sigmoid function, with the dimension of s being 1 x 1 x 1 x C.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \tag{3.3}$$

In this equation, the sigmoid function is denoted by σ, the ReLU activation function by δ, and the weights assigned to each channel by s. Z represents the outcome of the squeezing process.

Ultimately, the weight is multiplied by the original feature's weight using the following formula:

$$s_c = F_{scale}(u_c, s_c) = u_c \cdot s_c \tag{3.4}$$

The number of channels is indicated by the numbers above each bar. The input size is reflected in the sheer numbers on the left and right.

3.5 Comparison between Channel Attention Module (CAM) vs. Squeeze-and-Excitation (SE) Module

Both the Channel Attention Module (CAM) and the Squeeze-and-Excitation (SE) module are techniques designed to improve the representation power of neural networks by focusing on important features.

Channel Attention Module (CAM)

1. The purpose of CAM aims to enhance feature maps by identifying and emphasizing 'what' features are important across different channels.
2. CAM mechanisms are:
    - **Aggregation:** CAM aggregates spatial information using pooling operations to create a summary representation of each channel. This is usually done by applying average-pooling and max-pooling operations along the spatial dimensions, producing two 1D feature maps.
    - **Combination:** The two pooled feature maps are concatenated or added.
    - **Weight Generation:** A fully connected layer or a series of layers are used to generate a set of weights for each channel based on the aggregated feature maps.
    - **Attention:** These weights are applied to the original feature map, scaling the channels to emphasize important features.

3. The key Characteristics of CAM are:
    - **Pooling Operations:** Utilizes average and max pooling to summarize spatial information.
    - **Convolutional Layers:** May use fully connected layers for generating attention weights.
    - **Output:** Produces channel-wise attention weights which modulate the feature map.
4. The strengths of CAM are:
    - Directly focuses on the importance of each channel.
    - Often simpler to implement and integrate with existing architectures.

Squeeze-and-Excitation (SE) Module

1. The purpose of SE module aims to improve the representational power of a network by explicitly modeling the interdependencies between channels.
2. SE mechanisms are:
    - **Squeeze:** The SE module first performs global average pooling across the spatial dimensions to generate a channel descriptor. This operation condenses the spatial information into a single value per channel.
    - **Excitation:** The condensed channel descriptor is then passed through a series of fully connected layers (often a bottleneck architecture with reduction and expansion) to learn channel-wise dependencies. This typically involves a ReLU activation followed by a sigmoid function.
    - **Scaling:** The resulting weights (excitation weights) are applied to the original feature map through channel-wise multiplication, modulating the importance of each channel.
3. The key Characteristics of SE module are:
    - **Global Average Pooling:** Reduces each channel to a single representative value.
    - **Fully Connected Layers:** Uses a bottleneck structure to capture dependencies and generate attention weights.
    - **Output:** Produces channel-wise scaling factors applied multiplicatively to the feature map.
4. The strengths of SE module are:
    - Captures complex interdependencies between channels through the bottleneck structure.
    - Proven effectiveness in improving performance across various computer vision tasks.

Aspect	Channel Attention Module (CAM)	Squeeze-and-Excitation (SE) Module
Focus	Importance of each channel	Interdependencies between channels
Pooling Method	Average and max pooling	Global average pooling
Weight Generation:	Fully connected layers	Fully connected layers with bottleneck
Complexity	Moderate	Moderate to High
Application	Emphasizes important channels	Models complex channel interactions
Performance Impact	Enhances feature representation	Significant improvements in various tasks
Integration	Easily integrated	Easily integrated, slightly more complex

Table 3.1: Comparison between CAM and SE modules

Both CAM and SE modules are effective in enhancing neural network performance by focusing on important features. CAM provides a straightforward way to emphasize important channels, while SE modules offer a more sophisticated approach by modeling channel interdependencies. The choice between them depends on the specific requirements of the task and the complexity one is willing to manage in the model architecture. Table 3.1 shows the comparison between Channel Attention Module (CAM) and Squeeze-and-Excitation (SE) Module.

3.6 Advantages of the Squeeze-and-Excitation (SE) Module over the Convolutional Attention Module (CAM)

The Squeeze-and-Excitation (SE) module and the Convolutional Attention Module (CAM) both aim to enhance the representational power of neural networks, but they do so using different mechanisms. Here are some advantages of the SE module over CAM:

1. **Simplicity and Efficiency:**
   - **SE Module:** The SE module is conceptually simpler and computationally more efficient. It uses global average pooling followed by two fully connected layers to learn channel-wise dependencies. This straightforward mechanism reduces the complexity of the model, leading to faster computations and easier implementation.
   - **CAM:** In contrast, CAM combines both channel and spatial attention mechanisms, often requiring multiple pooling operations and convolutions, which can increase computational complexity and resource requirements.

2. **Focused Attention Mechanism:**
    - **SE Module:** By focusing solely on channel attention, the SE module effectively models inter-channel dependencies. This targeted approach ensures that the most informative channels are emphasized, which can be particularly beneficial in scenarios where channel importance is critical.
    - **CAM:** While CAM provides a dual attention mechanism (both channel and spatial), this can sometimes dilute the focus and add unnecessary complexity, especially if spatial attention is not crucial for the specific application.

3. **Lower Computational Overhead:**
    - **SE Module:** The computational overhead of the SE module is relatively low because it involves global average pooling and a few fully connected layers. This makes it suitable for applications with limited computational resources or for use in larger networks where efficiency is paramount.
    - **CAM:** The CAM, with its additional spatial attention mechanism, typically requires more operations, such as extra convolutions and pooling, leading to higher computational overhead.

4. **Ease of Integration:**
    - **SE Module:** The SE module can be easily integrated into existing network architectures without significant modifications. Its simplicity and modular nature make it an attractive option for enhancing network performance across various architectures and tasks.
    - **CAM:** Integrating CAM into existing networks can be more challenging due to its dual attention mechanism, requiring careful tuning and adaptation to ensure compatibility and optimal performance.

5. **Empirical Performance:**
    - **SE Module:** The SE module has demonstrated significant improvements in empirical performance across various computer vision tasks and network architectures, such as ResNet, Inception, and DenseNet. Its effectiveness and robustness have been validated in numerous studies and competitions.
    - **CAM:** While CAM also enhances performance, the added complexity and computational cost may not always translate to proportionally higher gains in performance compared to the SE module.

While both the SE module and CAM have their respective strengths, the SE module offers advantages in simplicity, efficiency, computational overhead, ease of integration, and proven empirical performance. These benefits make the SE module an attractive choice for enhancing neural network architectures, especially when computational resources are limited or when ease of implementation is a priority.

3.7 Applications of Squeeze-and-Excitation (SE) Module

The Squeeze-and-Excitation (SE) module has proven to be a valuable enhancement for various neural network architectures, and its applications span a wide range of tasks in computer vision and beyond. Here are some notable applications:

1. **Image Classification:** SE modules have been integrated into popular image classification architectures such as ResNet, Inception, and DenseNet, leading to improved performance on benchmark datasets like ImageNet. The SE module's ability to model channel interdependencies helps these networks focus on the most informative features.

2. **Object Detection:** In object detection tasks, SE modules can enhance the feature extraction capabilities of backbone networks (e.g., Faster R-CNN, YOLO, SSD). By emphasizing relevant channels, SE modules help in detecting objects more accurately, particularly in complex scenes.

3. **Semantic Segmentation:** Semantic segmentation, which involves classifying each pixel in an image, benefits from the SE module's channel attention mechanism. Networks like U-Net and DeepLab can incorporate SE modules to better differentiate between different classes in an image, leading to more precise segmentation results.

4. **Image Super-Resolution:** In image super-resolution, where the goal is to enhance the resolution of an image, SE modules have been used to improve the performance of convolutional neural networks (e.g., EDSR, SRGAN). The SE module helps in focusing on important features that contribute to generating high-quality high-resolution images.

5. **Generative Adversarial Networks (GANs):** GAN architectures, such as those used for image synthesis and style transfer, can integrate SE modules to improve the quality of generated images. The channel-wise attention provided by SE modules allows GANs to generate more realistic and detailed images by emphasizing critical features.

6. **Video Analysis:** In video analysis tasks like action recognition and video classification, SE modules can be applied to 3D convolutional networks to enhance the temporal and spatial feature extraction. This leads to better performance in recognizing complex actions and events in video sequences.

7. **Medical Image Analysis:** Medical imaging tasks, such as tumor detection, organ segmentation, and disease classification, benefit significantly from the SE module's ability to enhance feature representation. By integrating SE modules into networks like V-Net and U-Net, more accurate and reliable medical image analysis can be achieved.

8. **Reinforcement Learning:** In reinforcement learning, SE modules can be used in convolutional neural networks that process visual inputs from environments. By emphasizing relevant visual features, SE modules help improve the performance of agents in tasks like navigation, game playing, and robotic control.

9. **Remote Sensing:** SE modules have applications in remote sensing for tasks such as land cover classification, object detection in satellite images, and environmental monitoring. The improved feature extraction capabilities provided by SE modules lead to better analysis and interpretation of remote sensing data.

10. **Natural Language Processing (NLP):** Although less common, SE modules can be adapted for certain NLP tasks where multi-dimensional feature interactions are important. For instance, in models that combine text and image data (e.g., image captioning, visual question answering), SE modules can help in emphasizing relevant features across different modalities.

The SE module's ability to enhance feature representation by focusing on important channels has led to its successful application across a diverse array of fields, significantly improving performance in many tasks.

3.8 Loss Function

Identifying minor irregularities within extensive images presents a frequent obstacle in both segmentation and image analysis. The predominance of readily classifiable instances in the training dataset greatly influences the calculation of the loss function in these imbalanced datasets. Typically, intricate examples, which are relatively uncommon and may challenge the network, are overlooked in research studies. Altering the learning paradigm or selecting a suitable loss function is one potential approach to tackle the issue of data mismatch.

In a study by Akil M. et al. [11], inspiration is drawn from the Occipito-Temporal pathway for the model. The CNN model employs diverse receptive field sizes across consecutive layers to recognize crucial objects within a scene. Addressing the issue of class imbalance is achieved by selecting equal patches for each image, and experiments are conducted utilizing a weighted Cross-Entropy loss function.

ERV-Net [22] adopts a computational network based on 3D ShuffleNetV2 as its encoder, which effectively reduces GPU memory usage while improving overall performance. To prevent degradation, a Res-decoder with residual blocks is employed. Addressing convergence challenges and data imbalance, a fusion loss function combining dice loss and cross-entropy loss is utilized. This fusion approach helps mitigate the impact of unbalanced data in 3D segmentation tasks. Moreover, ERV-Net's output is normalized using the softmax function before applying the Dice loss. Experimental analysis explores the effectiveness of smooth cross-entropy loss, offering insights into managing class imbalance.

3.8.1 Cross Entropy Function

Cross Entropy is a fundamental and widely adopted loss function in semantic segmentation. It provides a robust framework for evaluating and optimizing the performance of segmentation models by quantifying the alignment between predicted and true class distributions. Its effectiveness in handling multi-class problems and compatibility with gradient-based optimization make it a cornerstone in the field of image segmentation. However, its sensitivity to class imbalance necessitates careful consideration and potential adjustments to ensure balanced performance across all classes.The formula for Cross Entropy is given by:

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1-p), & \text{otherwise} \end{cases} \tag{3.5}$$

Despite its widespread effectiveness, Cross Entropy loss has notable limitations, especially when dealing with imbalanced class distributions common in pixel classification tasks. Using Cross Entropy as the loss function in such scenarios results in training each pixel with an absolute value, making it difficult to achieve optimal performance due to the dominance of more frequent classes. This imbalance can lead to suboptimal learning, as the model might become biased towards the majority class, neglecting the minority classes. To overcome these challenges, we propose the Weighted Focal Loss Function, which is designed to address the issues posed by imbalanced class distributions. This advanced loss function builds on the original Focal Loss, which mitigates class imbalance by down-weighting the loss assigned to well-classified examples, thereby focusing more on hard-to-classify, minority-class examples.

The Weighted Focal Loss incorporates class-specific weights to further address class imbalance, making it particularly suitable for segmentation tasks with significant class frequency disparities. By assigning higher weights to minority classes and using a focusing parameter, this loss function ensures that the model learns more effectively from underrepresented classes, improves the model's ability to distinguish fine-grained classes, and reduces bias towards the majority class. This enhanced approach is particularly beneficial in applications like medical image segmentation, where the accurate detection of small pathological regions is crucial. The Weighted Focal Loss Function thus represents a significant advancement over traditional Cross Entropy loss, offering a more robust and balanced method for training models in pixel-level classification tasks.

3.9 Performance Metrics

Evaluating the performance of a U-Net model on the BraTS 2020 segmentation dataset using the dice coefficient and Intersection over Union (IoU) metrics is a standard and insightful approach in medical image segmentation tasks. These metrics provide a comprehensive assessment of the model's accuracy in delineating the boundaries and shapes of brain tumors

within medical images.

Both the Dice Coefficient and IoU are essential for evaluating the segmentation performance of medical images, as they emphasize different aspects of accuracy. The Dice Coefficient is more sensitive to the presence of small tumor regions, making it useful for detecting minor but clinically significant areas. In contrast, IoU provides a stricter measure of overall accuracy by considering the union of predicted and actual segments, thus penalizing both over-segmentation and under-segmentation.

Applying these metrics to the BraTS 2020 dataset helps in rigorously assessing the U-Net model's ability to segment brain tumors accurately. Accurate segmentation is critical for clinical applications, such as planning surgeries, monitoring treatment response, and predicting patient outcomes. By utilizing the Dice Coefficient and IoU, researchers and clinicians can ensure that the segmentation models are reliable and capable of capturing the complex and heterogeneous nature of brain tumors, leading to better patient care and treatment planning.

3.9.1 Intersection over Union (IoU)

The Intersection Over Union (IoU), also known as the Jaccard Index, is a crucial metric used to evaluate the performance of deep learning algorithms, especially in image segmentation and object detection tasks. This metric helps estimate how well a predicted mask or bounding box matches the ground truth data.

To calculate IoU, first need to determine the area of overlap between the predicted segmentation or bounding box and the ground truth. This overlap represents the common area that both the predicted and actual segments share. Next, calculate the area of union, which is the total area covered by both the predicted and ground truth segments combined. Figure 3.8 depicts the evaluation of Intersection over Union (IoU).

The IoU score is then obtained by dividing the area of overlap by the area of union. The formula for IoU is:

$$\text{Intersection over Union (IoU)} = \frac{Areaof Overlap}{Areaof Union} \tag{3.6}$$

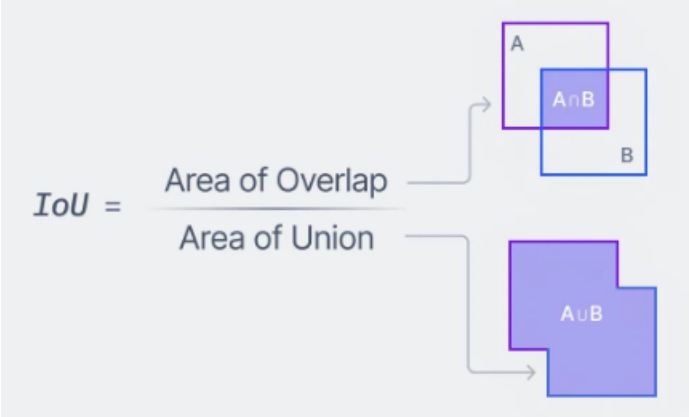

Figure 3.10: Intersection over Union

This score ranges from 0 to 1, where a score of 1 indicates a perfect match between the predicted and ground truth segments, and a score of 0 indicates no overlap at all. A higher IoU score means better performance of the deep learning model, as it suggests that the predicted segmentation closely aligns with the actual segmentation.

For example, if a deep learning model predicts a tumor region in a medical image, the IoU metric can be used to measure how accurately the model's prediction matches the actual tumor region annotated by medical experts. This helps in determining the effectiveness of the model in identifying and segmenting important areas within the images.

Using IoU as an evaluation metric is beneficial because it provides a clear and intuitive measure of accuracy, taking into account both false positives (areas incorrectly predicted as part of the object) and false negatives (areas that are part of the object but missed by the prediction). This makes IoU a valuable tool for improving and validating deep learning models in various applications, including medical image analysis, autonomous driving, and more.

For binary classification, it is written as:

$$\text{Intersection over Union (IoU)} = \frac{TP}{TP + FP + FN} \tag{3.7}$$

3.9.2 Dice Coefficient

The Dice coefficient, also known as the F1-score, is a crucial metric for evaluating the performance of image segmentation models. It is particularly useful in medical image analysis,

such as the segmentation tasks in the BraTS 2020 dataset. The Dice coefficient measures the accuracy of the model by calculating the harmonic mean of precision and recall.

To compute the Dice coefficient, follow these steps:

- **Intersection Calculation:** Determine the number of pixels that overlap between the predicted segmentation and the ground truth segmentation. This represents the intersection of the two sets.

- **Total Pixels Calculation:** Add up the total number of pixels in both the predicted segmentation and the ground truth segmentation. This gives you the union of the two sets.

- **Dice Coefficient Formula:** The Dice coefficient is calculated using the formula:

$$\text{Dice Coefficient } = \frac{2 * Intersection}{TotalNumberofPixelsinBothImages} \tag{3.8}$$

This formula effectively doubles the weight of the intersection, emphasizing the importance of correctly predicted pixels.

For example, if a model predicts a tumor region in a brain MRI, the Dice coefficient helps evaluate how well the predicted region matches the actual tumor annotated by medical professionals. A higher Dice coefficient indicates better performance, meaning the predicted segmentation is more accurate.

Figure 3.9 illustrates the evaluation of the Dice coefficient. It visually compares the predicted segmentation with the ground truth, highlighting the overlap (intersection) and the total pixels in both images. This visual representation aids in understanding how well the model performs and where improvements might be needed.

Using the Dice coefficient as a metric is beneficial because it balances both false positives (incorrectly predicted pixels) and false negatives (missed pixels), providing a comprehensive measure of segmentation accuracy. This makes it an essential tool for assessing and improving deep learning models in tasks requiring precise image analysis.

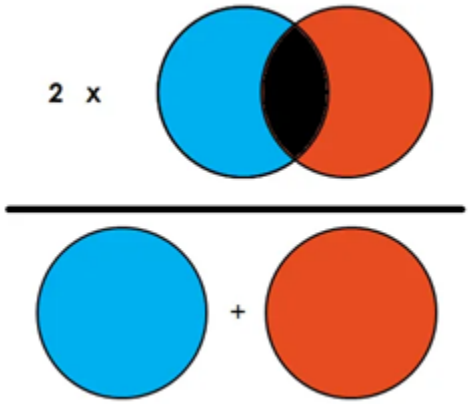

Figure 3.11: Dice Coefficient

The weighted average of precision and recall is the F1-score. It considers both false positives and false negatives to determine the model's overall accuracy.

$$\text{Dice Coefficient } = 2 * \frac{\text{Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{3.9}$$

Chapter 4

Experimental Analysis and Results

4.1 Experiments on data pre-processing task

This section describes the data preparation steps. Due to the varying sizes of images in the training dataset, they needed to be resized before being used as input for the model. The resolution of input images was standardized to 128 x 128 x 128 pixels. Rectangular images were resized so that their shortest side was 128 pixels, and then they were cropped to include only the central 128 x 128 x 128 square. It is important to note that the network requires input images to be in the 128 x 128 x 128 shape.

The BraTS 2020 Segmentation Dataset, available on Kaggle, offers 3D brain MRIs with voxel dimensions of 128 x 128 x 128. Each case includes four MRI modalities from MRI examinations, including Post-contrast T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). To utilize the most relevant data, the Native (T1) modality is omitted, and a stacked input is formed by combining T1ce, T2, and T2-FLAIR modalities, leveraging their complementary information.

The dataset also provides a ground truth segmentation map, classifying each voxel into one of four categories: Unlabeled volume, Necrotic and non-enhancing tumor core (NCR/NET), Peritumoral edema (ED), and GD-enhancing tumor (ET). All imaging datasets have been manually segmented and validated by experienced neuro-radiologists.

4.2 Environmental Setup

The experiments using the BraTS2020 Dataset on a 3D attention-based U-Net model for multi-region brain tumor segmentation were conducted on Google Colab, a cloud-based Python IDE. The setup included a GPU with the following specifications: 1x Tesla K80, compute capability 3.7, 2496 CUDA cores, and 12GB GDDR5 VRAM.

For the training process, the ADAM optimizer was utilized in conjunction with the categorical cross-entropy loss function. The choice of this optimizer and loss function was aimed at ensuring efficient convergence and effective handling of the multi-class segmentation task inherent in the BraTS2020 Dataset.

Performance analysis was carried out using accuracy and the Dice Coefficient as evaluation metrics. These metrics were chosen to provide a comprehensive assessment of the model's ability to correctly segment different regions of brain tumors. The Dice Coefficient,

in particular, is crucial for evaluating the overlap between the predicted and ground truth segmentations, making it an essential metric for segmentation tasks.

Through this rigorous evaluation, the best-performing model was identified, showcasing the effectiveness of the 3D attention-based U-Net in accurately segmenting multi-region brain tumors.

4.3 Results

From the training, validation, and testing results of accuracy and Dice Coefficient, as shown in Table 4.1, it is evident that the proposed model achieves the best results on the BraTS 2020 dataset in terms of both accuracy and Dice Coefficient. This indicates the model's superior performance in segmenting brain tumor regions compared to other models evaluated.

The model was trained for 30 epochs, with each epoch consisting of batches of 32 images. This batch size was chosen to balance memory efficiency and training speed. The learning rate was set to 0.0001, a value selected to ensure stable convergence of the model during training.

We employed the Adam optimization method, which is known for its adaptive learning rate capabilities and efficiency in handling large datasets. For the activation function in the convolutional layers, we used ReLU (Rectified Linear Unit). ReLU is preferred for its ability to introduce non-linearity and help the network learn complex patterns.

The evaluation of the model was carried out using two key metrics: accuracy and the Dice Coefficient. Accuracy measures the overall correctness of the segmentation, while the Dice Coefficient evaluates the overlap between the predicted and ground truth segmentations, providing a more specific measure of segmentation performance.

Figure 4.2 presents the accuracy curves for both the training and validation sets. These curves illustrate how the model's accuracy improved over the course of the 30 epochs. The training accuracy curve shows the model's performance on the training data, indicating how well the model learns from the provided dataset. The validation accuracy curve shows the model's performance on unseen data, reflecting its generalization capability.

Through careful tuning of the learning rate, optimization method, and activation function, the proposed model was able to achieve excellent performance metrics. The results underscore the model's effectiveness in accurately segmenting multi-region brain tumors, making it a valuable tool for medical image analysis.

Table 4.1: Performance scores of the BraTS 2020 dataset

	Accuracy	**Dice Coefficient**		
		ET	**TC**	**WT**
Training	0.995	0.82	0.84	0.86
Validation	0.992	0.83	0.89	0.87
Testing	0.990	0.84	0.86	0.89

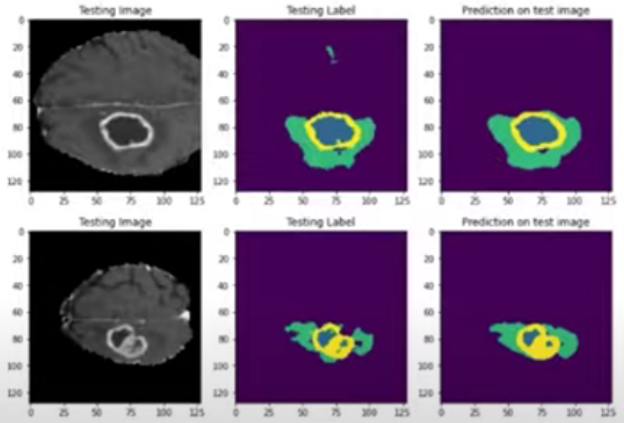

Figure 4.1: segmentation result of BraTS 2020

Figure 4.1 provides visual results, displaying input images alongside their corresponding masks. In these images, different regions of the tumors are highlighted in specific colors to indicate various tumor components.

The green areas in the projected images represent the tumor core regions (TC). These are the central parts of the tumor that include the active, non-necrotic tumor cells. The accurate identification of these regions is crucial for assessing the severity and extent of the tumor.

The yellow areas in the projected images represent the whole tumor regions (WT). These encompass the entire visible tumor, including the active tumor cells, the necrotic core, and the surrounding edema. Highlighting these areas helps in understanding the full extent of the tumor's impact on the brain.

The blue areas in the projected images represent the enhancing tumor regions (ET). These regions are particularly important as they indicate areas where the tumor is actively growing and taking up contrast agent, showing the aggressive parts of the tumor.

This visual representation confirms that our model accurately identifies and delineates the different tumor regions. By comparing the input images with the highlighted masks, it is evident that the model effectively distinguishes between the tumor core, the whole tumor, and the enhancing tumor regions. This accuracy in segmentation demonstrates the model's effectiveness in providing detailed and clinically relevant information, which is essential for diagnosis, treatment planning, and monitoring of brain tumors.

The clarity and precision of these visual results validate the model's performance and reinforce its potential utility in medical image analysis, specifically for multi-region brain tumor segmentation.

4.4 Comparison: proposed model versus state-of-the-art works with the same dataset

Table 2 presents a comprehensive comparison between the proposed model and various published works focused on BraTS 2020 dataset. This comparison highlights the strengths and advantages of the proposed model over existing methods.

Table 4.2: Comparison: proposed model versus state-of-the-art works with the same dataset

Author	Method	**Dice Coefficient**		
		ET	**TC**	**WT**
Mora et al. [15]	3D U-Net + V-Net	0.77	0.81	0.82
Qamar et al. [16]	HI-Net	0.79	0.87	0.83
Ahmad et al. [17]	3D U-Net	0.71	0.88	0.75
Proposed	**Proposed Model**	**0.89**	**0.84**	**0.86**

Mora et al. [15] utilized a basic 3U-Net model and a V-net model for brain tumor segmentation. However, these models were prone to overfitting due to a high number of parameters. Overfitting can lead to poor generalization on new, unseen data, making these models less robust.

Qamar et al. [16] employed a HI-Net CNN model, which proved effective for brain tumor segmentation. However, the HI-Net model required lengthy training times, making it a significant time investment. The extended training periods could be a drawback in scenarios where quick model deployment and iteration are needed.

Ahmad et al. [17] utilized a 3D U-Net architecture for brain tumor segmentation. While their model showed promising results, the proposed model demonstrated superior performance in several key areas. Specifically, the proposed model excels in the semantic segmentation of brain tumors, effectively handling a substantial volume of images and generating precise pixel-level segmentation maps for each class.

The proposed model's advanced segmentation capabilities are evidenced by the evaluation metrics, particularly the multi-class Dice scores. These scores confirm that the proposed model outperformed the earlier studies, showcasing its proficiency in accurately segmenting brain tumors. The model's robustness and precision in generating detailed segmentation maps for each tumor class underscore its potential as a valuable tool in medical image analysis.

Overall, the proposed model not only addresses the limitations of previous works but also sets a new benchmark in the semantic segmentation of brain tumors. Its ability to manage large datasets efficiently and produce accurate results makes it a significant advancement in the field of brain tumor segmentation.

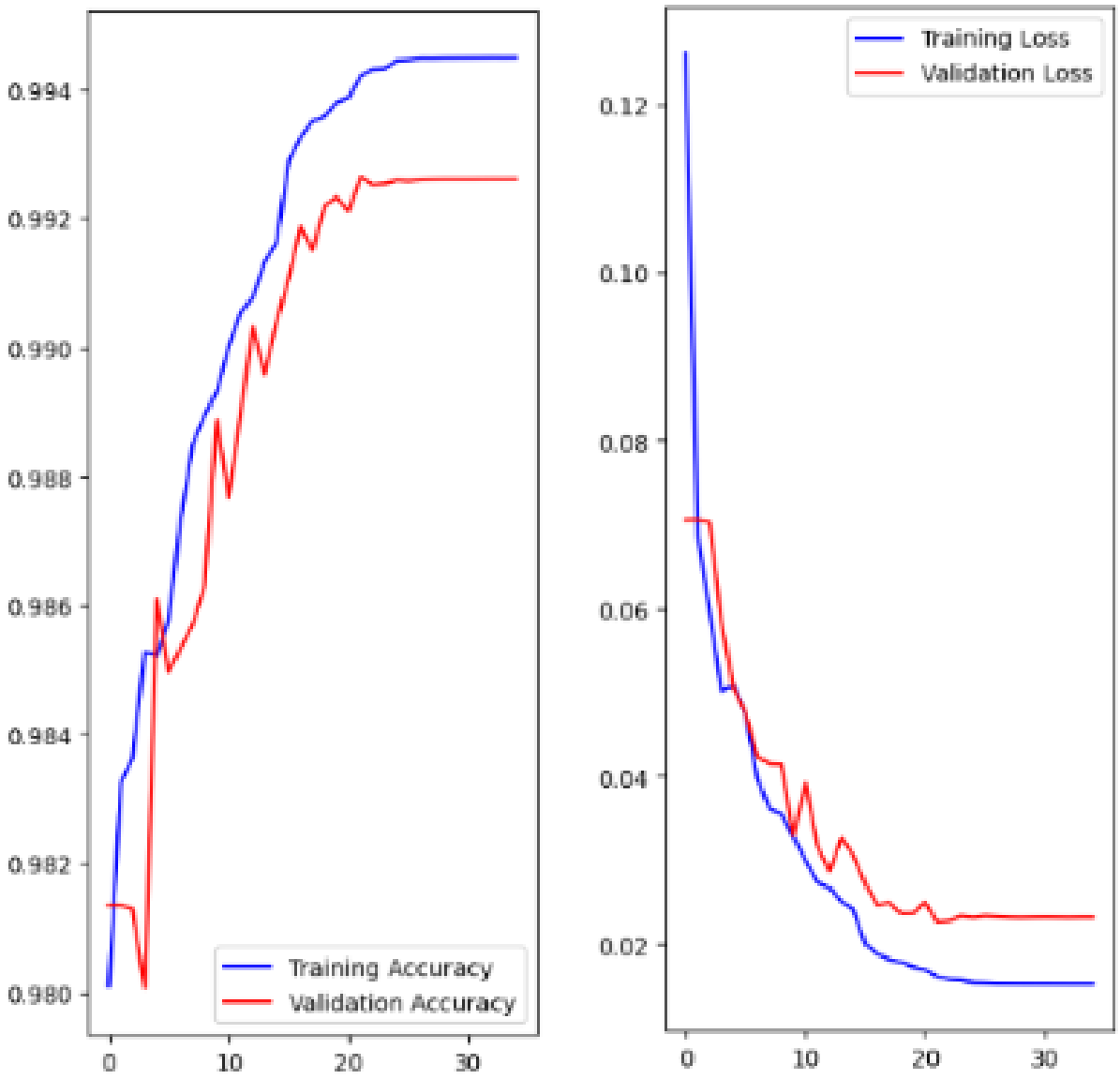

Figure 4.2: Accuracy - Loss graph of BraTS 2020 dataset

Chapter 5

Conclusion and Future Scope

The proposed methodology leverages a hybrid architecture, combining the Squeeze-and-Excitation (SE) module as the encoder and U-Net as the decoder, to achieve precise segmentation of brain tumors in MRI images. This hybrid approach uses the strengths of both architectures to improve segmentation accuracy.

A novel 3D attention-based U-Net is employed for the segmentation task. The segmentation performance of the model is analyzed using evaluation metrics such as the Dice coefficient. The Dice coefficient measures the overlap between the predicted segmentation and the ground truth, providing a reliable metric for model performance.

The proposed model achieves better results compared to existing methods. It exhibits remarkable precision, yielding Dice coefficients of 0.84 for the Tumor Core (TC), 0.89 for the Whole Tumor (WT), and 0.86 for the Enhancing Tumor (ET) regions. These high Dice coefficients underscore the model's efficacy in accurately delineating brain tumors.

This model approach not only enhances the understanding of tumor segmentation but also offers a promising avenue for advancing the field of medical image analysis. By accurately segmenting different regions of brain tumors, the model provides valuable information for diagnosis and treatment planning.

Furthermore, this comprehensive approach improves our understanding of tumor segmentation and offers a significant advancement in MRI-based brain tumor detection and localization. The hybrid architecture and the novel 3D attention-based U-Net together create a powerful tool for medical professionals, enhancing the precision and reliability of brain tumor segmentation in MRI images.

References

[1] R. Ayachi and N. Ben Amor, "Brain tumor segmentation using support vector machines," in *European conference on symbolic and quantitative approaches to reasoning and uncertainty*. Springer, 2009, pp. 736–747.

[2] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using u-net based fully convolutional networks," in *Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings 21*. Springer, 2017, pp. 506–517.

[3] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[4] X. Feng, N. J. Tustison, S. H. Patel, and C. H. Meyer, "Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features," *Frontiers in computational neuroscience*, vol. 14, p. 25, 2020.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2016, pp. 179–187.

[7] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[8] S. Cui, L. Mao, J. Jiang, C. Liu, S. Xiong *et al.*, "Automatic semantic segmentation of brain gliomas from mri images using a deep cascaded neural network," *Journal of healthcare engineering*, vol. 2018, 2018.

[9] G.-C. Lin, W.-J. Wang, C.-M. Wang, and S.-Y. Sun, "Automated classification of multi-spectral mr images using linear discriminant analysis," *Computerized Medical Imaging and Graphics*, vol. 34, no. 4, pp. 251–268, 2010.

[10] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical image analysis*, vol. 43, pp. 98–111, 2018.

[11] Z. Akkus, I. Ali, J. Sedlar, T. L. Kline, J. P. Agrawal, I. F. Parney, C. Giannini, and B. J. Erickson, "Predicting 1p19q chromosomal deletion of low-grade gliomas from mr images using deep learning," *arXiv preprint arXiv:1611.06939*, 2016.