AMERICAN UNIVERSITY OF ARMENIA

College of Science and Engineering

BS in Data Science Bachelor Thesis

Exploring the Impact of Parameters on the Performance of Brain Tumor Segmentation Algorithms in the BraTS Challenge

Author:

Ela Khachatryan

Supervisor:

Varduhi Yeghiazaryan

May 9, 2024

Abstract

Accurate segmentation of brain tumors is crucial for correct diagnosis and treatment planning. U-Net segmentation is one of the most successful algorithms in medical image analysis. It has been in the list of top solutions of the BraTS benchmark. This paper does an in-depth analysis of a specific variation of 3D U-Net algorithm with slight modifications of the algorithm's parameters, namely batch size and training–test data quantity ratio. The data splitted into training and test with ratio 8:2 and batch size 2 (instead of 1) slightly outperformed the original source algorithm's result. This is because the model has more data to learn from and train on. Also, when training with batch size 2 and concatenating MRI 3D images, the model can see some general patterns he could not observe with batch size 1.

Contents

1. Introduction

In medical image analysis, where every detail is essential in accurately diagnosing the patient, the Multimodal Brain Tumor Segmentation (BraTS) challenge is one of the field's cornerstones [1]. The BraTS is an annual challenge to improve automated brain tumor segmentation using multimodal magnetic resonance imaging (MRI) data. It was first initiated in 2012, and since it has become one of the standards in medical imaging analysis, attracting participants worldwide. Participants are provided with the same medical data and are challenged to develop algorithms with maximum accuracy for diagnosing brain tumors and their degrees.

Another crucial objective of the challenge is promoting collaboration and knowledge exchange between researchers worldwide. This challenge brings together people from diverse backgrounds, such as researchers, clinicians, and data scientists, who, collaborating and sharing their work, can eventually reach solutions to complex problems and improve the state of the art in medical image analysis.

Over the years, the BraTS challenge has encompassed different tasks. These tasks included brain tumor segmentation in Multiparametric Magnetic Resonance Imaging(mpMRI) scans, prediction of the Methylguanine-DNA-methyltransferase(MGMT) promoter methylation status in mpMRI scans, prediction of patient overall survival, evaluation of algorithmic uncertainty in tumor segmentations, and much more [2]. This report only concentrates on the task of brain tumor segmentation in mpMRI scans.

The brain tumor segmentation task includes the segmentation of brain tumors using multimodal MRI scans and experts provided clinically acquired data. Algorithms are implemented to have maximally improved performances and attain machine-generated diagnoses with minimum

error. The algorithms divide the tumor area into various glioma sub-regions: "enhancing tumor" (ET), the "tumor core" (TC), and the "whole tumor" (WT). Participants must upload their segmentation methods using a standardized method and receive a rank based on standard evaluation metrics—Dice Score, Hausdorff distance [3].

The BraTS challenge has multiple essential purposes. The BraTS challenge's primary purpose is to improve medical imaging analysis in brain tumor segmentation by benchmarking the performance of different segmentation algorithms. The challenge enables fair comparison between various methods since it provides standardized datasets and evaluation metrics. This way, it encourages the development of innovative algorithms that can accurately segment brain tumors using MRI scans. The final goal is to reach improved diagnostic accuracy and better segmentation outcomes. Another purpose the BraTS challenge follows is to enhance the treatment process of brain tumors. Generally, the diagnosis of a brain tumor is very time-and resource-consuming. By getting automated, accurate diagnoses using the algorithms derived from the BraTS challenge, field professionals can spend their time- and resources on the treatment rather than the diagnosis. Moreover, since the best expert board-certified neuroradiologists label the ground truth data used in the BraTS challenge [1], the challenge can enable more accurate patient treatment planning. The doctors can better understand tumors' size, shape, and location, allowing them to choose the best-suiting treatment for the given patient.

While hundreds of algorithms have been suggested since the BraTS benchmark first started in 2012, it's important to understand what factors cause improvements/deterioration of the algorithm's accuracy and how it can be enhanced. In this paper, several such factors influencing an algorithm's performance will be discussed.

2. Literature Review

2.1 Brain Tumor Segmentation Challenges

The main reason behind the initialization of the BraTS challenge was the need for a uniform evaluation system for the best computer algorithms in medical image analysis [1]. The BraTS challenge has had many different tasks during its existence. When starting in 2012, the BraTS challenge had one primary goal—to track the current state-of-the-art multimodal automated brain tumor segmentation and compare different algorithmic approaches [4]. Over time, the benchmark became more comprehensive and included more tasks, where the results of the brain tumor segmentation were used to enable additional research [5]. In 2016, the participants were asked not only to estimate the size and location of the tumor but also to predict whether the tumor area is "progressing", "shrinking", or stable [6]. Gradually, the tasks of the challenge became more complex—in 2017 and 2018, the participants had two distinct tasks—implement segmentation of gliomas in pre-operative multimodal Magnetic Resonance Image (MRI) scans and predict patient overall survival (OS) from pre-operative scans [7] [8]. BraTS 2019 and 2020 extended to experimentally evaluating the uncertainty in tumor segmentations [2], [9]. Lately, BraTS 2023 and BraTS 2024 include a variety of essential tasks—including the BraTS Challenge on Relevant Augmentation Techniques, BraTS Adult Glioma Post Treatment Challenge, and ASNR-MICCAI BraTS MRI Synthesis Challenge (BraSyn) [10] [11]. However, through time, the main objective of the benchmark remains to identify the current state-of-the-art segmentation algorithms for brain gliomas.

2.2 BraTS Dataset and Its Evolution Since 2012

The BraTS challenge data has continued growing in size since 2012. In 2012-13, the BraTS challenge was initiated with around 60 Magnetic Resonance Image(MRI) scans. The data included both clinical and synthetic records. The records were also diverse in glioma grades(it had high-grade and low-grade glioma patients) [1]. The data quantity was rather small in 2012-2013, however it started growing in 2014-2016 and almost doubled in quantity in 2017. Validation sets were also added in 2017.(Table 2.1)

Clinical data included pre- and post-therapy mpMRI scans from patients collected over multiple years, utilizing various MR scanners operating at 1.5T and 3T field strengths, employing diverse scanning protocols including 2D and 3D imaging techniques [12]. For the synthetic data, MR images imitating clinical records were generated using TumorSim software, combining physical and statistical models to simulate the four subregions of the tumor. The algorithm used anatomical maps from healthy subjects and introduced variations such as noise and intensity inhomogeneities to simulate diverse imaging conditions [1]. Although synthetic data already came with ground truth annotations, the clinical data must be thoroughly examined and labeled. In order to be able to annotate different visual structures, a special annotation protocol was established. Usually, 1-4 domain professionals examined and manually segmented the data. A 15-year expertise single board-certified neuro-radiologist reviewed the annotations for maximal accuracy and consistency [5].

Four types of MRI volumes were decided to be used for segmenting the clinical data. These four types of MRI modalities include [1]:

T1: T1-weighted image, with 1–6 mm slice thickness;

T1c: T1-weighted, contrasted Gadolinium(T1GD) image;

T2: T2-weighted image, with 2–6 mm slice thickness;

T2-FLAIR: T2-weighted FLAIR image, 2–6 mm slice thickness.

Each of these modalities are useful for identifying a different subregion.

However, it has been shown that the most useful ones in annotations were T1GD and T2-FLAIR [5].

Another requirement of the annotation protocol was that all modalities' volumes were to be co-registered to the T1c MRI and resampled to 1mm isotropic resolution [12]. Also, all scans were skull-striped to guarantee anonymization of the patients [1].

The tumor subregions defined in the protocol were designed to be imaged-based rather than biological structures [1]. During 2012-2016 the annotations were being made for four subregions—"edema"(ED), "non-enhancing(solid) core/tumor"(NET), "necrotic(or fluid- filled) core"(NCR), and "enhancing core" or Active Tumor(AT)(Fig. 2.1).

ED describes the peritumoral edematous and invaded tissue that is fairly easily defined on the T2-weighted images. NET are parts of the high-grade tumor do not enhance, but they are clearly distinguishable from the surrounding vasogenic edema. These are the parts used for identifying low grade gliomas(LGG)(since LGG are quite hard to determine). NET as well is best identified using T2. NCR describes the necrotic core, and is often cystic. AT describes the enhancing regions within the gross tumor abnormality, but not the necrotic center [5].

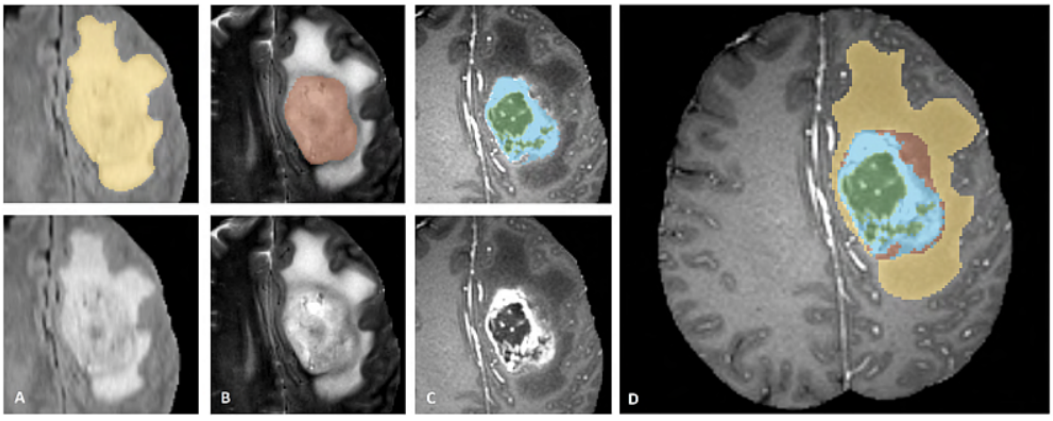

Figure 2.1: Tumor appearance on three imaging modalities (A = T2-FLAIR, B = T2, C = T1c) with manual annotations, and fusion of the three labels on the right (D). From left to right: whole tumor (yellow), tumor core (red), enhancing tumor structures (light blue), surrounding the cystic/necrotic components of the core (green). (Reproduced from [12].)

Later in 2017 BraTS challenge changed the annotation protocol to

use three subregions—Whole tumor(WT), Tumor core(TC), and Active Tumor(AT) [5]. The WT was the union of ED, NET, NCR and AT, and was best discoverable by T2-Flair. The TC was the union of everything except the edema, and was best identified using T1GD and T1. AT covers only the enhancing tumor. [5] [12].

The BraTS dataset has increased heavily over time in size(Table 2.1). More clinical data annotated by experts have been added to BraTS datasets. The main objective of the BraTS is to keep this benchmark an open sourced and continously developing challenge for the scientific community [5].

Year	Total data	Training data	Validation data	Testing data	Tasks	Type of data
2012	50	35	N/A	15	Segmentation	Pre-operative only
2013	60	35	N/A	25		
2014	238	200	N/A	38	(Segmentation Disease progression)	Longitudinal
2015	253	200	N/A	53		
2016	391	200	N/A	191		
2017	477	285	46	146	Segmentation	Pre-operative only
2018	542	285	66	191		

Table 2.1: BraTS Data 2012-2018 Summary ([5])

2.3 Evaluation Metrics

The Multimodal Brain Tumor Segmentation Challenge 2020 focused on detecting tumor sub-regions (ET, TC, WT), predicting patient overall survival from pre-operative MRI scans, and evaluating uncertainty in segmentation. The primary evaluation metrics included the Dice Similarity Coefficient (DSC), which measures overlap by calculating

$$\frac{2\times TP}{2\times TP+FP+FN},$$

where TP, FP, and FN represent the number of true positive, false positive, and false negative voxels, respectively. This metric, insensitive to the background's extent, complements the 95th percentile of the Hausdorff Distance (HD95), detecting maximum contour discrepancies and penalizing even small outlier errors that could significantly impact clinical outcomes. Additional metrics used were sensitivity TP/(TP+FN) and specificity TN/(TN+FP), providing a comprehensive evaluation of the segmentation's accuracy on the tumor sub-regions ET, TC, and WT [3], [13].

2.4 Common Approaches and Algorithms

When the BraTS challenge first landed in 2012, there were only a few participants, and the results were far from accurate. The number of the participants and the algorithms increased heavily since then [12].

The algorithms used for brain tumor segmentation usually fall into two groups—generative and discriminative. The generative algorithms use domain expert information in order to attain automated segmentations [1]. Pathology, MRI physics, and radiology knowledge are necessary for correctly implementing image analysis based on intensity and shape. First, experts identify the regions with the expected features of a healthy brain. In this case, since MRI images are being used, the process involves determining the intensity of healthy classes. After defining what intensity features are considered to be normal, tumors and edema are treated as intensity abnormalities or outliers. Afterward, unsupervised clustering techniques are used to determine normal tissues and abnormalities. Lastly, spatial and geometric properties are used to specify the abnormal tissues' location correctly [14]. However, the generative models have one big limitation—the difficulty of transforming semantic image feature interpretations into probabilistic models [1]. The discriminative approach is the other widespread approach that also handles the limitations present in generative models. The discriminative models do not require prior domain knowledge. They directly learn the feature differences between abnormalities and healthy tissues from manually labeled images. The annotated images later become the ground truth tables for the training process. Participants need significant amounts of labeled data to obtain robust and precise results. First, algorithms extract the features for each voxel from anatomical maps such as MRI images. Then, the algorithms use these features to implement supervised learning processes(classification) that return the labeled segmentation maps as outputs per each input image. One limitation of such algorithms is the annotation protocol used for the training data that needs to be followed by each input image [1].
One possible solution which avoids the limitations of both discriminative

and generative models is the focus on joint nerative-discriminative techniques. This approach involves generative component's probabilistic models in the preprocessing step and discriminative model's supervised algorithms in the main training step [1].

2.4.1 The BraTS Algorithms Through the Years

As mentioned above, in 2012, when the BraTS challenge debuted, the algorithms' evaluation scores were relatively low(averaging 0.14-0.70 for the whole tumor and 0.09-0.37 for the tumor core). 8 out of 12 algorithms used were fully automated. The trends for the year were Random Forests (RF) classification, Markov Random Field (MRF), and logistic regression. In 2012, most of the best performances were observed at Random Forests [12]. Random forests are ensembles of randomly different decision trees. Training each decision tree involves adjusting the parameters of the split function at every node to maximize information gain when dividing the training data. The testing is executed by pushing each feature vector through the tree models and testing each split node until a leaf node is achieved. The algorithm calculates labels by averaging the posteriors of the leave nodes of all trees. One advantage of random forests is that they can naturally manage multi-class problems and deliver a probabilistic output instead of hard-label separations [15].
Compared to 2012, the evaluation score results were much higher in 2013 (average Dice score range of 0.71-0.87 for WT, 0.46-0.78 for TC, and 0.52-0.74 for ET). Inspired by the success of Random Forests, most participants in 2013 continued the utilization of these algorithms for the BraTS challenge. 4 participants out of 10 used RF algorithms, and 3 of them were considered as top solutions. The best-ranked solution was a concatenation of Random Forest models. During the concatenation, the output of one model served as an input for the next RF. Binary morphological processing was used as the final step for the advancement of the labeling results [12].

Summarizing the results of 2012 and 2013, two general patterns are

observed. First, even though some models performed exceptionally well for the time, the inter-rater agreement among expert clinicians still showed better results. Second, despite the accuracy of some individual models, the concatenated algorithms showed much better performance. They were the state-of-the-art algorithms during 2012 and 2013 [5]. (Fig. 2.2).

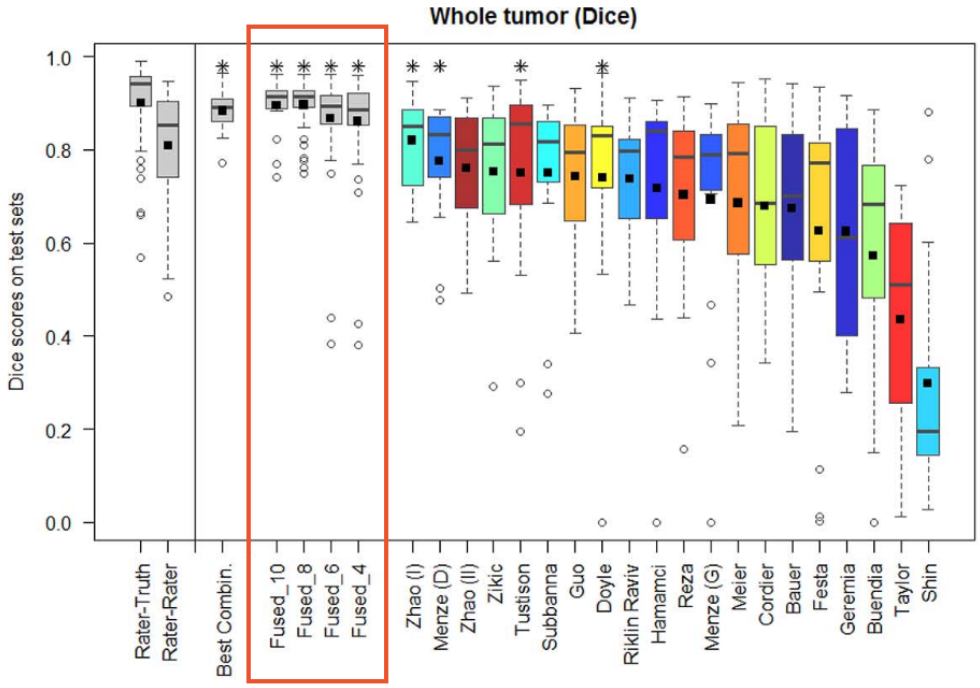

Figure 2.2: Summary results of the BraTS 2012-2013. Label fusion (red outline) out-performs all individual methods and the inter-rater agreement. (Reproduced from [1].)

Moving forward to 2014-2016, the scientific community observed a new tendency in algorithms. Though 4 out of 8 submissions for BraTS 2014 challenge were based on RF-s, Convolutional Neural Networks(CNN) were introduced then and gained popularity in the field [12].

The CNN uses 3D patches extracted from MRI scans, each representing a small cube of voxels. Each patch is centered around the voxel to be classified, capturing local spatial information in three dimensions (x, y, z) along with the channel data. The algorithm pushes these patches through multiple layers of 3D convolutional filters. Each convolution layer applies a set of learned filters and a non-linear activation function, progressively reducing the spatial dimension but extracting relevant features. The final

convolution layer is equipped with as many filters as there are target classes (e.g., different tumor types and normal tissue), each filter producing a spatial map of predictions [16].

In the BraTS 2017 and 2018 challenges, the tendency among the algorithms submitted was the use of convolutional neural networks (CNNs), specifically, the majority used U-net inspired models [12].
The U-net network architecture is illustrated in Figure 2.3. It consists of an encoder (left side) and a decoder (right side). The encoder mainly reduces the quality of the picture by applying 3×3 convolutions, followed by a rectified linear unit (ReLU), and a 2×2 max pooling operation. The decoder restores the image to its original dimensions by up-sampling. It does upsampling by applying 2×2 convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU [17]. U-Net, in contrast to traditional CNN-s, uses feature concatenation from the encoder and decoder to restore the original size of the image. [18].

Among the submissions, a significant number utilized CNN architectures, with 16 models based on U-Net or its variant, V-Net. These models were broadly used due to their effectiveness in handling volumetric data, which is crucial for medical image segmentation.

The tendency towards the U-Net and its variations remained after 2017–2018 as well. The winner algorithm of BraTS 2020 was a very simple implementation of U-Net, called nnU-Net [19]. First, nnU-Net preprocesses the input image voxels by normalizing them. The nnU-Net follows a 3D U-Net-like pattern. It has an encoder and a decoder that are connected by skip connections. nnU-Net relies on plain convolutions for feature extraction. Dice loss and cross-entropy loss are used for the training. The model was trained with a batch size of 2. The model won the BraTS 2020 benchmark with Dice scores of 0.88, 0.85 and 0.82 and HD95 values of 8.498, 17.337 and 17.805 for whole tumor, tumor core and enhancing tumor, respectively [19].

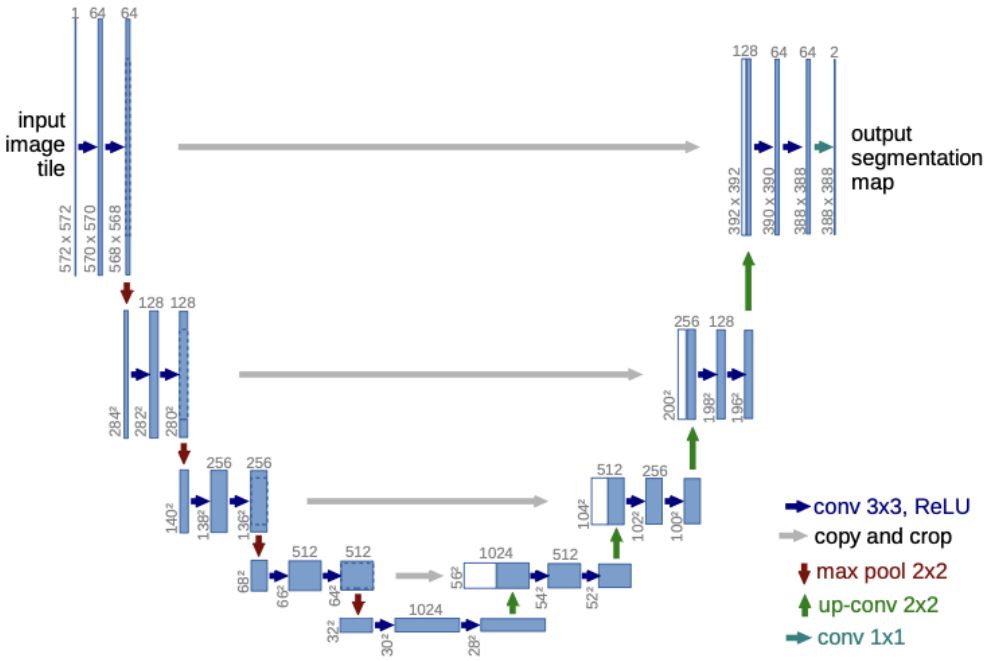

Figure 2.3: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. (Reproduced from [17].)

Another successful algorithm that was in the top 10 of BraTS 2020 competition was "Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-net neural networks". The model again works based on a simple U-Net with minor modifications. The model's architecture can be found in Figure 2.4. In contrast to nnU-Net, this model trains its data only using Dice loss and does its encoding in 3 stages. The model achieved a Dice of 0.79, 0.89 and 0.84, and Hausdorff (95%) of 20.4, 6.7 and 19.5mm on the final test dataset [3].

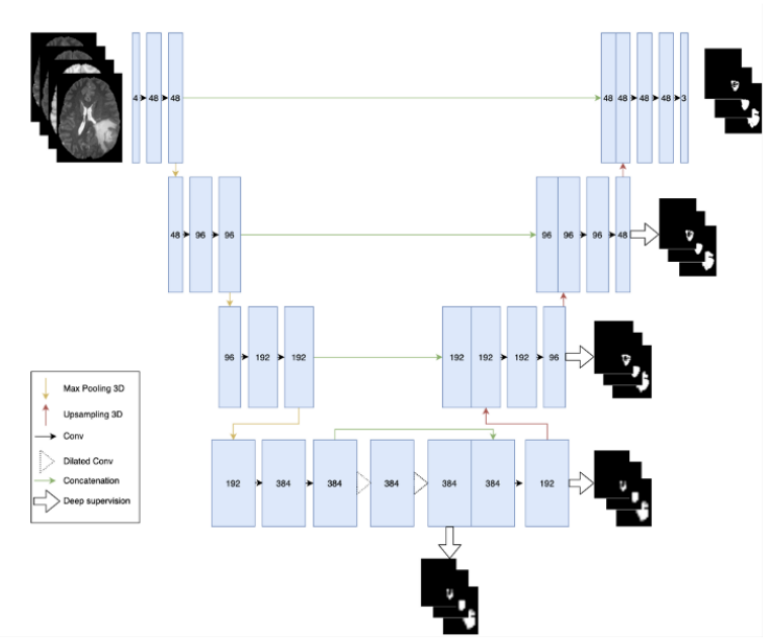

Figure 2.4: Neural Network Architecture: 3D U-Net [35] with minor modifications. (Reproduced from [3].)

3. Contributions

While researching brain tumor segmentation, I explored many articles discussing different approaches and algorithms in this field. I decided to pick an algorithm and experiment with it. My goal was to experiment by adjusting some parameters within this algorithm. This experiment would allow me to see how changing factors could impact its effectiveness. Through this process, I aimed to gain a more profound understanding of the algorithm's functionality and explore possibilities for improving its performance by changing its parameters.
Throughout my research, I observed a tendency in methodologies over the years towards using deep learning techniques. In Particular, 3D U-net architecture has shown one of the most successful results. For this reason, I focused on an algorithm called "Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-net neural networks"(discussed in **??**) for my investigation. This algorithm was a BraTS 2020 challenge solution, ranked among the top ten teams [3].

3.1 Data Preprocessing

Officially, registration in the challenge and data requests are needed to access the authorized datasets of the BraTS challenges. Consequently, due to the shortage of time, direct access to the official BraTS 2020 dataset was not available during the course of this study. Therefore, I attained the equivalent dataset for BraTS 2020 from Kaggle, a widely recognized platform for data science competitions and datasets. The participants did not have access to the official testing data and validation ground truth labels, so the latter was missing in the available dataset. I resolved this problem by generating an artificial test dataset from the existing training data. I developed a script that shuffles the training dataset and randomly

splits it into training and testing subsets according to a widely used ratio—80% training dataset and 20% testing dataset.
. This setup prepares the data for further experiments.

3.2 Evaluation Scores

Since the participants did not have access to the testing data, they did not include functionality to calculate or record the experiment's evaluation metrics. Working with an artificially generated testing dataset, I found these evaluation scores essential for understanding the model's performance and assessing the outcomes of my future experiments. I integrated a new functionality into the source inference file functions that calculates all four key evaluation metrics for the predicted segmentations. This advancement not only computes individual metrics but also averages them to identify broader performance patterns. By doing so, it simplifies the process of comparing these metrics across different experimental setups. I also added a small script in a separate notebook file calculating averages per each region—the ET, TC, and WT in order to compare the results region-wise as well.

3.3 Batch Size Modifications

The first parameter I decided to experiment with is the batch size. The source model was initially configured to train with a batch size of 1. Sometimes, a larger batch size can reach more precise estimates of the gradients and potentially better optimum minima—thus resulting in more accurate predictions [20]. Considering this potential improvement and the available GPU hardware capacities, I considered training my network by doubling the batch size to 2. Finally, it was intended to compare the results with those obtained from the original batch size setting. Though setting a higher batch size made the algorithm execute much faster and showed better accuracy, the GPU limitations did not allow me to go beyond a batch

size of 2 and take this experiment further.

3.4 Splitting Ratio Modifications

Constrained by the GPU limitations that prevented continuing experiments with larger batch sizes, I came up with the idea of experimenting with split sizes of training and testing datasets. The ratio between training and testing datasets can potentially improve an algorithm's performance. As the testing dataset was generated artificially, I had the privilege of making different versions with different ratios and comparing the outcomes.
Consequently, I generated datasets with training-to-testing ratios of 9:1, 8:2, 7:3, and 6:4. For each configuration, I prepared the data, conducted the split, and proceeded with the testing and inference phases to save and compare the results.
Lastly, I tried integrating a data augmentation model into the source code. I completed the data preprocessing phase and was focusing on standardizing the sizes of the MRI slices. but due to the time constraints this will be left as a future work.

4. Results and discussion

4.1 Results

The evaluation scores for the source data were as following: Hausdorff (95%) of 20.6, 5.7, 4.3 mm; Dice of 0.81, 0.85 and 0.91 for the enhancing tumor, tumor core and whole tumor, respectively[3].

As discussed in the Contributions(Chapter 3), in this work the inference has been executed in 5 different setups – the original split(split with ratio 8:2) with batch size 1, the original split with batch size 2, the split with ratio 6:4 with batch size 1, the split with ratio 7:3 with batch size 1, and the split with ratio 9:1 with batch size 1.
As indicated in Table 4.1, the results of the first setup are different from the initial source's results since the algorithms have been trained on different datasets.

Table 4.1: Summary of Evaluation Metrics for Inference with Training-Testing Ratio 8:2 and Batch 1

Metric	Hausdorff	Dice	Sensitivity	Specificity
Global Mean	11.64	0.85	0.89	0.999
ET averages	8.68	0.80	0.87	0.999
TC averages	8.35	0.86	0.90	0.999
WT averages	17.77	0.89	0.90	0.999

Comparing the results between Tables 4.1 and Table 4.2, we can clearly indicate that in scope of this experiment the Specificity stays the same, Sensitivity is also relatively stable. The Dice scores modify slightly in specific regions, but still taken globally they are the same. The only noticable difference is the Hausdorff distance, which has decreased when using batch 2 in all regions—taking the global average Hausdorff distance from 11.64 to 6.70. This is a good indicator—stating that a small improvement is alredy observed when using batch size of 2.

Table 4.2: Summary of Evaluation Metrics for Inference with Training-Testing Ratio 8:2 and Batch 2

Metric	Hausdorff	Dice	Sensitivity	Specificity
Global Mean	6.70	0.85	0.89	0.999
ET averages	7.14	0.79	0.87	0.999
TC averages	6.67	0.87	0.90	0.999
WT averages	15.19	0.88	0.89	0.999

Table 4.3: Summary of Evaluation Metrics for Inference with Training-Testing Ratio 6:4 and Batch 1

Metric	Hausdorff	Dice	Sensitivity	Specificity
Global Mean	15.94	0.83	0.87	0.999
ET averages	10.33	0.79	0.80	0.999
TC averages	9.70	0.84	0.87	0.999
WT averages	27.42	0.86	0.94	0.997

Moving forward and comparing the evaluation metrics for predictions trained on data split with Training-Testing Ratio 6:4(Table 4.3) to the predictions trained on data split with Training-Testing Ratio 7:3(Table 4.4), we can see that Specificity again stays the same, Sensitivity slightly increases in the case of Training-Testing Ratio 7:3, which is a good sign, Dice score stays relatively the same, and Hausdorff distance decreases, which indicates that the setup with Training-Testing Ratio 7:3 outputs better predictions than the one with Training-Testing Ratio 6:4.

Table 4.4: Summary of Evaluation Metrics for Inference with Training-Testing Ratio 7:3 and Batch 1

Metric	Hausdorff	Dice	Sensitivity	Specificity
Global Mean	13.97	0.83	0.91	0.999
ET averages	11.45	0.76	0.88	0.999
TC averages	10.42	0.90	0.88	0.999
WT averages	19.89	0.88	0.95	0.998

Comparing the results between Tables 4.4 and Table 4.1, we can see that Specificity stays the same in setups of Training-Testing Ratio 7:3 and 8:2, the Sensitivity is a bit lower when the training data is 80%, Dice scores increases by 0.2 points and Hausdorff decreases, which indicates that the setup with

80% training data works better than that with 70% training data.

Table 4.5: Summary of Evaluation Metrics for Inference with Training-Testing Ratio 9:1 and Batch 1

Metric	Hausdorff	Dice	Sensitivity	Specificity
Global Mean	14.09	0.82	0.88	0.999
ET averages	12.38	0.74	0.85	0.999
TC averages	9.32	0.84	0.85	0.999
WT averages	20.42	0.89	0.93	0.999

Finally, comparing the results between Tables 4.1 and Table 4.5 we can see that the predictions trained on data split with Training-Testing Ratio 8:2 still show better results(smaller Hausdorff distances and higher Dice scores) than the ones trained on data split with Training-Testing Ratio 8:2. However, the difference is very small(around 2.37 for Hausdorff distance and 0.02 for Dice score).

4.2 Discussion and Future Work

Finalizing the results observed in Section 4.1, the model with parameter modifications slightly surpassed the original source model. We can see a general trend (with one exception) that more training data results in better segmentation results that are shown with higher Dice scores and lower Hausdorff distances. Note that the batch size was fixed to 1 for comparing the split effect. The metrics get better as we move from predictions of inference with a training-test ratio of 6:4 to inference with a training-test ratio of 8:2. The only exception is the results getting slightly worse in the case of a training-test ratio of 9:1 compared to the ratio of 8:2. Overall, we can state that the more training data resulted in better evaluation metrics in the scope of this experiment. This is because the model has more data to learn from and train on.

Another observation made is that changing the batch size from 1 to 2 also improved the model's performance in terms of accuracy. This can result from more precise estimates of the gradients that batch size 2 can cause. Also, when training with batch size 2 and concatenating 3D MRI images,

the model can catch some patterns it could not observe before.

To achieve better results, several steps could lead to improvement. First, it would be great to have access to the original dataset and do the training on the 100% of the training dataset.

Second, implementing *K*-fold cross-validation could enhance our segmentation model's robustness and generalizability. Due to time constraints, as each training session lasted for over several hours, this approach was not feasible within the scope of the current study. Exploring this method could provide more comprehensive insights into the model's performance across different subsets of data.

Lastly, data augmentation could be very helpful by artificially increasing the amount of data. The current experiment does not include any data augmentation since it was not present in the basic setup of the source model architecture. However, I discovered that a complex data augmentation model was used in the winner architecture of the same year BraTS 2020 challenge [19]. I tried integrating the data augmentation model into the "Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-net neural networks" algorithm. I finished the data preprocessing part and was working on making the MRI slice sizes consistent. However, it was not included in the scope of this project due to time constraints and the complexity of the algorithms.

5. Conclusion

This paper describes a model proposed in the BraTS 2020 Challenge with modified parameters and its most relevant results , modified to enhance the results by adjusting the parameters. The experiments included modifying the algorithm to train the data using batch size 2 instead of batch size 1; experimenting with splitting the data into training and testing datasets using different ratios to catch general patterns on how data splitting ratios can effect model performance. I showed that there are slight improvements when executing the training phase on a bigger dataset. Also, the results were better when executing the code with batch size 2 rather than with batch size 1. The model's final best results consisted of Dices of 0.79, 0.87, 0.88 , and 0.55 validation accuracy , and Hausdorff distances of 7.14, 6.67, 15.19 for Enhancing Tumor Tumor Core and Whole Tumor, respectively. Note that those results were achieved with a 3D U-Net model with modified batch size, random split training, and testing data of the ratio 8:2, respectively.

Bibliography

[1] 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.

[2] *Brain tumor segmentation (brats) challenge 2020*, https://www.med.upenn.edu/cbica/brats-2019/, Accessed: April 30, 2024.

[3] T. Henry, A. Carré, M. Lerousseau, *et al.*, "Brain tumor segmentation with self-ensembled, deeply-supervised 3d u-net neural networks: A brats 2020 challenge solution," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*, Springer, 2021, pp. 327–339.

[4] M. Society, "Proceedings of the miccai-brats 2012 challenge," in *Proceedings of the MICCAI-BRATS 2012 Challenge*, MICCAI Society, 2012. [Online]. Available: https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2012_proceedings.pdf.

[5] 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.

[6] M. Society, "Proceedings of the miccai-brats 2016 challenge," in *Proceedings of the MICCAI-BRATS 2016 Challenge*, MICCAI Society, 2016. [Online]. Available: https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2016_proceedings.pdf.

[7] M. Society, "Proceedings of the miccai-brats 2017 challenge," in *Proceedings of the MICCAI-BRATS 2017 Challenge*, MICCAI Society, 2017. [Online]. Available: https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf.

[8] M. Society, "Proceedings of the miccai-brats 2018 challenge," in *Proceedings of the MICCAI-BRATS 2018 Challenge*, MICCAI Society, 2018. [Online]. Available: https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2018_proceedings_shortPapers.pdf.

[9] *Brain tumor segmentation (brats) challenge 2020*, https://www.med.upenn.edu/cbica/brats2020/, Accessed: April 30, 2024.

[10] Synapse, *Brats challenges 2023*, https://www.synapse.org/#!Synapse:syn51156910/wiki/, Accessed: April 30, 2024, 2023.

[11] Synapse, *Brats challenges 2024*, https://www.synapse.org/#!Synapse:syn53708249/wiki/, Accessed: April 30, 2024, 2024.

[12] M. Ghaffari, A. Sowmya, and R. Oliver, "Automated brain tumor segmentation using multimodal brain scans: A survey based on models submitted to the brats 2012–2018 challenges," *IEEE reviews in biomedical engineering*, vol. 13, pp. 156–168, 2019.

[13] V. Sundaresan, L. Griffanti, and M. Jenkinson, "Brain tumour segmentation using a triplanar ensemble of u-nets on mr images," in *International MICCAI brainlesion workshop*, Springer, 2020, pp. 340–353.

[14] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig, "A brain tumor segmentation framework based on outlier detection," *Medical image analysis*, vol. 8, no. 3, pp. 275–283, 2004.

[15] S. Bauer, T. Fejes, J. Slotboom, R. Wiest, L.-P. Nolte, and M. Reyes, "Segmentation of brain tumor images based on integrated hierarchical classification and regularization," in *MICCAI BraTS Workshop. Nice: Miccai Society*, vol. 11, 2012.

[16] M. Society, "Multi- modal brain tumor segmentation using deep convolutional neural networks," in *Proceedings of the MICCAI-BRATS 2014 Challenge*, MICCAI Society, 2014, pp. 31–35. [Online]. Available: https://people.csail.mit.edu/menze/papers/proceedings_miccai_brats_2014.pdf.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[18] R. Preetha, M. J. P. Priyadarsini, and J. Nisha, "Comparative study on architecture of deep neural networks for segmentation of brain tumor using magnetic resonance images," *IEEE Access*, 2023.

[19] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[20] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT express*, vol. 6, no. 4, pp. 312–315, 2020.