Image Segmentation for Vestibular Schwannoma

Bo Zhang
Center of Data Science, NYU
bz854@nyu.edu

Chengyu Chen
Center of Data Science, NYU
cc7027@nyu.edu

Xinyi Gu
Center of Data Science, NYU
xg2085@nyu.edu

December 16, 2021

Abstract

Although MRI technology provides strong supports for the diagnosis of brain diseases, the images could not self-interpret without human judgments. Our work achieves auto-segmentation that produces contours of the tumor area to facilitate physicians' and researchers' diagnosis and treatment planning before radiosurgery for Vestibular Schwannoma. We present a pipeline that takes 3D MRI data and predicts brain tumor masks using Convolutional Neural Networks. We also explored the possibility of applying this model to other brain diseases such as Metastatic brain tumors. Our work builds a strong foundation for a treatment follow-up system to achieve automatic detection of tumor regrowth or decay after the surgery based on volume change. Code source: https://github.com/NYUROAI/Capstone-VS.

1 Introduction

As an important tool for diagnosis, MRI provides strong support for more advanced brain tumor treatments, such as microsurgical resection and stereotactic radiosurgery with its high soft-tissue contrast. Vestibular Schwannoma is a benign tumor originating in the ear canal also known as an acoustic neuroma. The recommended treatment methods are primarily radiosurgery. This kind of tumor is located nearby cochlea and trigeminal nerve, and thus the tumor may cause problems such as loss of hearing and facial paralysis. Our research goal for this project is automating measuring growth rates of Vestibular Schwannoma before and after radiosurgery. Faster-growing tumors are both more likely to cause problems for the patient and respond better to radiosurgery than slower-growing tumors. In order to achieve this, we need to do auto-segmentation to separate the tumor from other brain tissues on brain MRIs, which are T1 weighted sequence of brain MRIs that is contrast-enhanced. We choose a convolutional neural network, in particular, U-Net shaped CNN, as our primary model since we are dealing with grid signals, and U-Net is widely accepted in MRI study. Our work will be based on previous segmentation work [Ellis and Aizenberg, 2021] from BraTS Challenge 2020. After fine-tuning, the model reaches a median dice score of 0.88, which indicates that 88% of tumor regions are correctly labeled. Also, as the model predicts the possibility of each voxel being a tumor, the model is not limited to single-tumor prediction.

2 Related Work

Deep learning becomes a popular topic in the healthcare field, and CNN, in particular, is now widely applied in MRI studies. There are several works that focus on different body parts like the liver, whole heart, and spine segmentation. Brain segmentation is one of the major fields, especially due to the hardness in brain disease treatment. The Multi-modal Brain Tumor Segmentation (BraTS) challenge sets an open environment for researchers to study how to perform such tasks in a more efficient and effective way. The challenge evaluates state-of-the-art methods for segmenting glioma-type brain tumors presented in an annual event that invites participants to train and test their segmentation methods on the BraTS dataset [Menze et al., 2015]. Previous BraTS challenges have shown that deep learning is one of the most accurate methods to segment tumor regions. For example, in BraTS 2017, [Wang et al., 2019] proposed a deep learning network using multiple layers of anisotropic and dilated convolutional filters, winning second place in that challenge. Recently, the U-Net convolutional neural network model gains popularity in the top-performing teams winning the challenge. In BraTS 2018, [Isensee et al., 2019] demonstrated that a generic U-Net architecture with a few minor modifications is enough to achieve competitive performance. Inspired by these challenges, our project applies the 3D U-Net convolutional neural network structure on 3D brain MR imaging provided by NYU Langone to segment the Vestibular Schwannoma brain tumor. In our problem setting, we only have a single weighted sequence of MRI (T1w) for the tumor while BraTS offers multiple sequences that are differently weighted. This way, T1w MRI alone might under-represent the contrast between tumor tissue and its surrounding brain tissues, making it harder for the model to detect the tumor region. By automating the tumor contouring process, the model can offer references of tumor positions and volumes to senior physicians at scale. Also, the model pipeline can be applied to other brain tumors such as Metastatic brain tumors and tumors on other body parts as only fine-tuning is needed for these downstream tasks.

3 Problem Definition and Algorithm

3.1 Task

The aim of this research project is to perform auto-segmentation of Vestibular Schwannoma brain tumor from MR imaging with performance comparable to manual segmentation, which will provide clinicians and researchers a reference in planning for radiosurgery. We managed to perform this image segmentation task with a U-Net convolutional neural network model. Specifically, the model takes the 3D grayscale MR imaging of the whole brain as inputs and outputs a 3D matrix that maps the tumor segmentation. In Figure 1, the left image is a slice of the 3D MRI along the z-axis, and the right image is the corresponding output tumor segmentation map of the same slice.

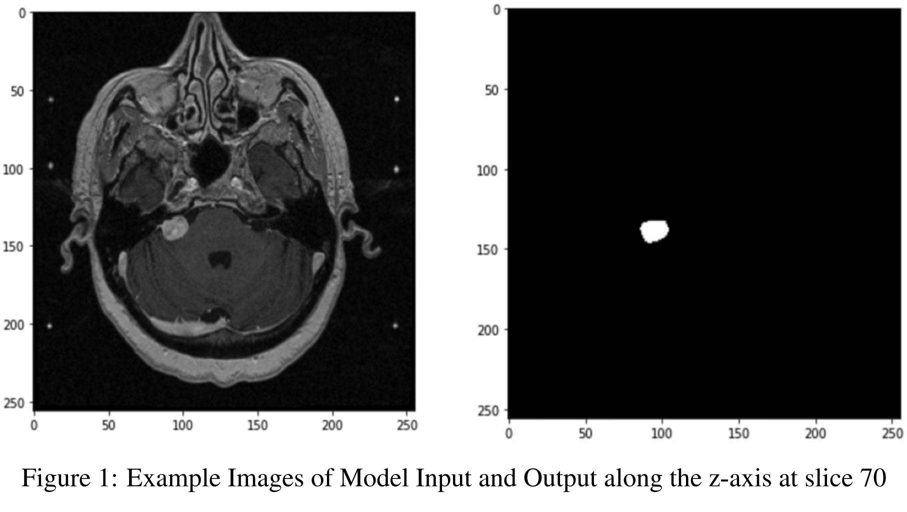


Figure 1: Example Images of Model Input and Output along the z-axis at slice 70

With a well-trained U-Net-CNN model, tumor masks will be generated as a whole in 20 seconds whereas hand-drawn masks need to be inspected slice by slice by physicians. However, when encountering heterogeneous tumors or other special cases, the prediction still needs human judgment to be involved. In dealing with those issues, we come up with another measurement to indicate model confidence. The segmentation map of tumors will serve as a reference along with the confidence score to supplement physicians' surgery planning. The outcome will then be fed to the volume calculating algorithm we developed when determining tumor response after treatment by measuring its volume change.

3.2 Algorithm

The model that we utilized in this task is a U-Net structured convolutional neural network. In general, this model architecture employs convolutional layers that encode the information stored in the images progressively rendering smaller resolutions in each step, and then decodes the output of the encoder along with residual connections from previous layers to restore the original resolution. The output will then be fed to a Sigmoid activation to obtain the probability scores.

The main idea behind CNN is to learn the feature mapping of an image and exploit it to make more nuanced feature mapping. In image segmentation, we not only need to convert the feature map into a vector but also to reconstruct an image (i.e. the tumor segmentation map) from this vector. Since we have already learned the feature mapping of the image while converting the image into a vector, we can use the same mapping to convert it back to the image. This is the main idea behind the U-Net structure. Figure 2 below demonstrated the U-Net Architecture in the image.

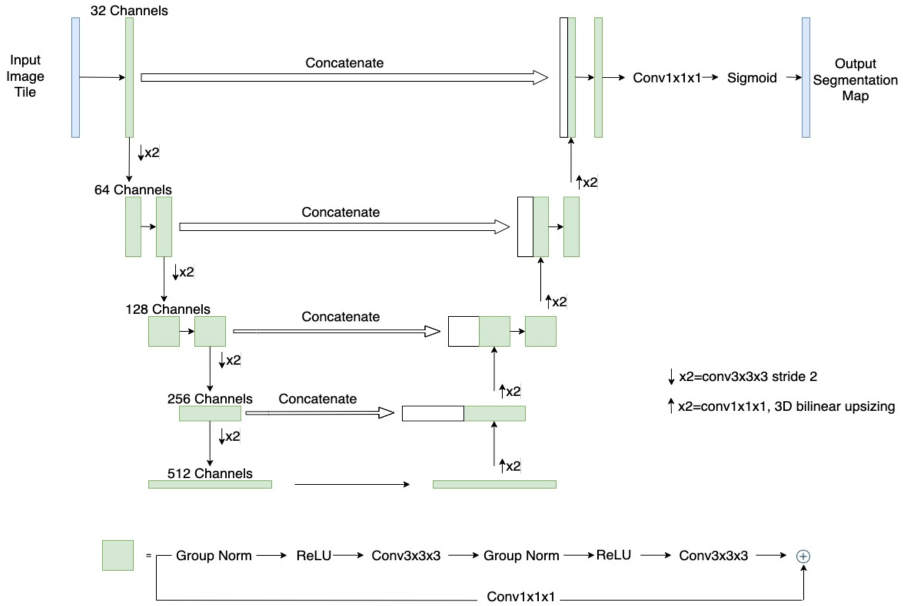


Figure 2: U-Net CNN Structure

The architecture consists of three sections: the contraction/encoding, the bottleneck, and the expansion/decoding sections. The encoding and decoding sections are highly symmetric and both have 5 layers. The encoding block starts from a base width of 32 channels, applies 3D filters in each layer, and ends with 512 channels which is the opposite of the decoding procedure. There are two residual blocks per layer. Each residual block consists of two convolutional blocks performing group normalization, followed by rectified linear unit activation and a $3 \times 3 \times 3$ convolution. In the expansion section, each time the input is also appended by feature maps of the corresponding contraction layer. This process guarantees that features learned while contracting the image will be used to reconstruct it. After that, a final $1 \times 1 \ \times 1$ convolution map linearly resamples the outputs from the 32 channels of the last decoding layer, and a sigmoid activation function is applied to predict the target segmentation map.

4 Experimental Evaluation

4.1 Data

The dataset is provided by NYU Langone, the department of Radiation Oncology of New York University. The dataset contains 3D MRIs where every single image is huge. So, we store our dataset in NYU Langone HPC (Big Purple) environment and utilize the GPU resources for computation. The entire set is constituted by brain MRIs from 356 de-identified patients. 53 of the patients are excluded out from the modeling stage. Those who experienced resections in previous treatments or those who had multiple brain diseases were excluded from the dataset as their tumors were heterogeneous compared to typical patients who have only Vestibular Schwannoma tumors. As a result, 303 patients remained as valid inputs.

For each patient, we are provided with a 3D scan of the whole brain and a 3D matrix that contains a hand-drawn mask of the Vestibular Schwannoma tumor area stored in Nifty files. The brain scan is in grayscale and contains values from 0 to 255 denoting different highlights for brain tissues (i.e larger the value, the brighter the tissue in the scan). On the other hand, the tumor mask only contains 0 and 255 where 255 indicates that the current pixel is within the tumor range. We convert the 255 values to 1 in the modeling setting during data processing.

The data does not come in uni-sized. For any given patient, the 3D matrices of brain scan and hand-drawn tumor mask have the same dimension and voxel size. However, the dimension and voxel size of the scan might be different from patient to patient. For older scans, the dimensions are $512 \times 512 \times x$ where x ranges from 76 to 82 with voxel size of $0.47 \times 0.47 \times 2.00$. As for the newer ones, the dimensions are $256 \times 256 \times 208$ with voxel size of $0.82 \times 0.82 \times 1.00$. To unify the dimensions while keeping the actual volume of the brain unchanged, we chose to down-sample the large scans. The down-sampling might lose some information contained in the original scan but would save computation time. With the automated pre-processing, future scans of various dimensions and sizes can be fed into the model even if the MRI scanner updates its image configuration. Thus, we developed a data processing pipeline that will calculate the actual size of the brain based on dimension and voxel size, and divide it by our target dimension to get a new voxel size. This new voxel size and the target dimension will become a reference of how we down-sample or up-sample the image.

In addition, data augmentation techniques are applied to the data at training time. For example, we add random Gaussian noise to the input images and blur the input images using a Gaussian kernel randomly with 50% probability per training iteration. Left-right mirroring and scale distortion are also applied. In summary, the augmentation process prevents the model from overfitting and thus makes the model more solid and general to the dataset.

4.2 Methodology

We split 303 patients into three sets, training, validation, and testing, with each of them containing 60%, 20%, and 20% of patients of the entire input set.

We will be using two measuring criteria, the dice loss, and confidence score. We also computed volumes of each tumor and compared the calculated result with ground-truth tumor volumes provided by our supervisors as compensation for the other two criteria. Its major task is to predict tumor progression before and after radiosurgery as a tracking measurement in future work.

Dice loss is the main criterion used during training and evaluation. It measures how well our predicted mask overlaps with the hand-drawn mask. However, when the future input comes in, where we do not have the ground truth hand-drawn tumor mask available, we design a confidence score to indicate how confident our model is on the predicted mask as an alternative.

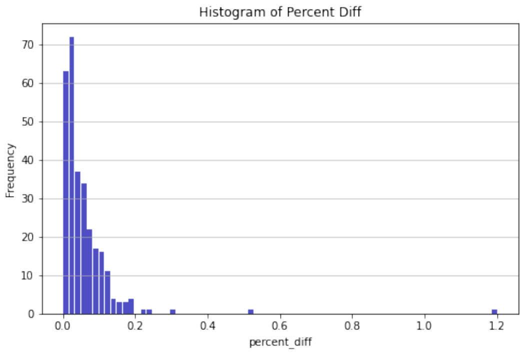


Figure 3: Percent Difference between Calculated and Ground-Truth Volume

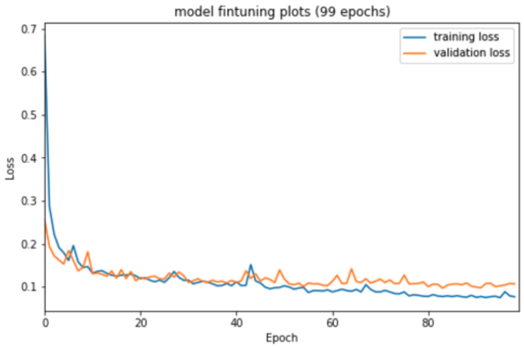


Figure 4: Learning Curve of the UNET Model

$$\text{Dice Score} = \frac{2\sum_{i=1}^{N} \hat{y}_i y_i}{\sum_{i=1}^{N} \hat{y_i}^2 + \sum_{i=1}^{N} y_i^2}$$

Volume Calculation: We developed an algorithm that calculates tumor volume based on pixel counts, which can also be used as a rough measurement for model performance. Because the tumor masks are stored in a 3D format, where each pixel has a thickness defined by the voxel size and tumors have irregular shapes, generating a contour line and using another mathematical way to calculate the volume will cause a large deviation. Thus, we choose to use the counting method to get tumor volumes by calculating the product of counts of the number of pixels and the volume of each pixel. This method, compared to other methods, has more flexibility because accuracy can be adjusted based on the dimension and voxel size of the image. The error rate is presented in the results session. If we want to obtain higher accuracy, we can up-sample the image to have a smaller voxel size for each pixel to obtain a more detailed counting.

Confidence Score: As discussed in the model architecture session, the last layer of our U-net model is a Sigmoid activation, and thus it turns outputs from the previous layer to values between 0 and 1. We consider this as a probability score of how likely each pixel is considered as a tumor pixel. At the final stage of the prediction, the algorithm decides whether a certain pixel is within the tumor area or not based on a threshold, which is chosen as 0.5. The confidence score is calculated as the average of probability scores within the tumor. We checked its validity by testing its correlation with dice loss using Pearson's R score. We hypothesize that the confidence score positively correlates with the dice score.

$$\text{Confidence Score} = \frac{1}{N}\sum_{i=1}^{N} prob_i, \text{ where } prob_i > 0.5$$

4.3 Results

Volume Calculation: As mentioned in the methodology part, we have developed an algorithm for tumor volume calculation based on tumor mask matrix. The performance is shown in the histogram(Figure3) below, where 95% of the patients has percent difference less than 15% and 60% of the patients has percent difference less than 5% .

Model Result: We applied the pre-trained model and fine-tuned it for extra 99 epochs on brain MRI data with manually labeled tumor segmentation maps using the High-Performance Computing environment provided by NYU Langone Health. The pre-trained model uses the Multi-modal Brain Tumor Segmentation(BraTS) dataset [Menze et al., 2015]: 3D brain MRI from 369 subjects with

Table 1: Performance on test set

Number of data	60
Mean	0.873023
Median	0.883270
Min	0.682197
Max	0.946937

T1w, contrast-enhanced T1w, T2w, and T2w-FLAIR images along with manually labeled tumor segmentation maps. Figure 4 demonstrated the learning curve of this fine-tuning procedure. The loss on both validation and training set decreases as the number of epochs increases. After 99 epochs, the dice loss achieves 0.07 on the training set(0.93 dice score) with validation loss achieving 0.1 (0.9 dice score). The training loss drops at a rapid speed within the first few epochs. The reason is that using the BraTs data, which have similar problem settings, the weights and parameters are already optimized. And thus we could obtain a fairly good result in fewer epochs to save computation.

Table 1 shows the model performance on the test set. There are prediction results with a total of 60 patients and the distribution of the dice score on the test set is right-skewed with the mean dice score at about 0.87 and median dice score at 0.88. Our model achieved similar accuracy with a single sequence (T1w) compared with the work done for BraTs challenge [Ellis and Aizenberg, 2021] which has a mean dice score at 0.9 when trained using 4 different weighted tumor scans (T1w, T2w, etc.)

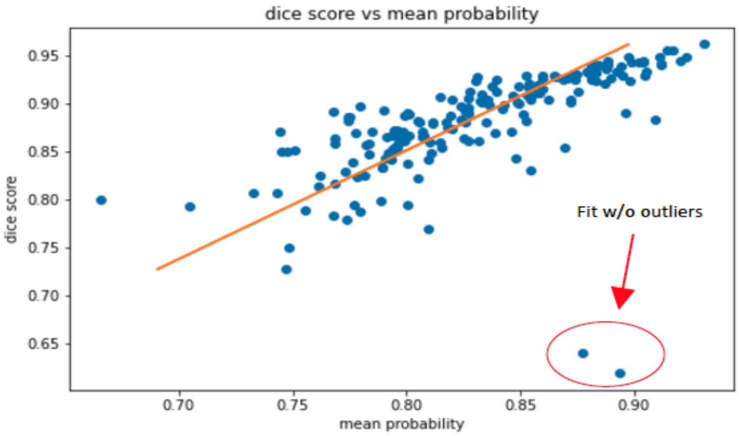
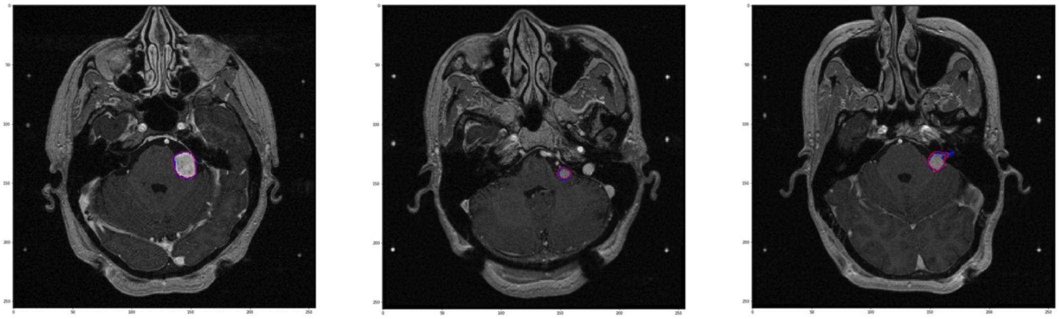


Figure 5: Confidence Score

Confidence Score: The model confidence score is an alternative to dice loss when a hand-drawn mask is not available for future inputs. As mentioned in the previous section, it is calculated as the mean of probability scores extracted from the Sigmoid activation output where pixels are selected if they pass the 0.5 thresholds. We used Pearson correlation to test the confidence score and dice score with a hypothesis that indicates probability and dice score don't have a linear relationship. The Pearson correlation is 0.64 and the p-value is $1.2e - 21$. With a p-value below 0.05, we reject the null hypothesis. We find that while mean probability increases, dice score also increases. Thus we verify the validity of using confidence score as a supplement for dice score.

In Figure 8, there are two outliers with high confidence scores but low dice scores. These two patients are special cases that should be left out when fitting into the model. One of them is from a patient after microsurgical resection where there is a dark region between brain stem. The other one is the case where the patient has two tumors on both sides of the brain. Thus, we can detect abnormal data points with the confidence score.

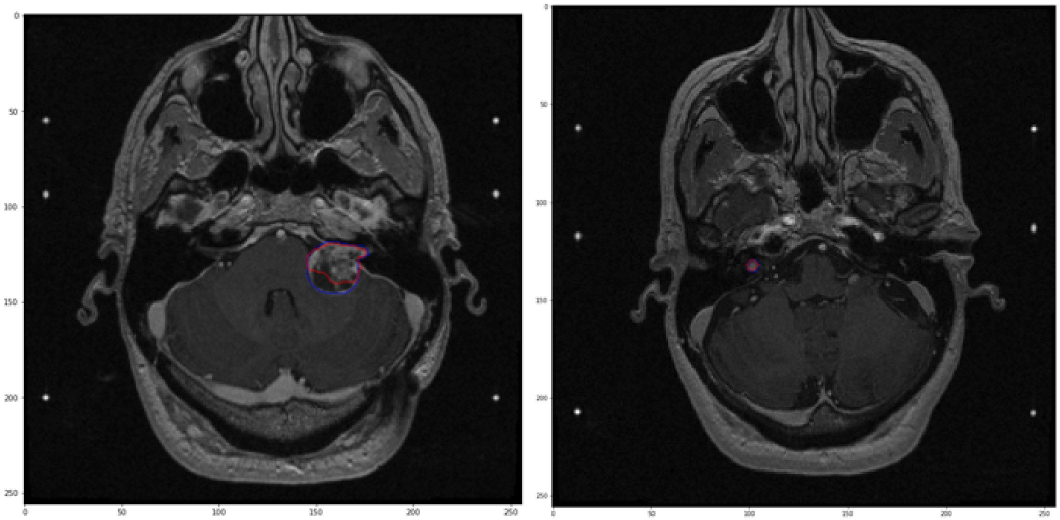


(a) True volume: 4883.9995 mm^3
Dice score: 0.947

(b) True volume: 1952.1194 mm^3
Dice score: 0.944

(c) True volume: 1022.1542 mm^3
Dice score: 0.897

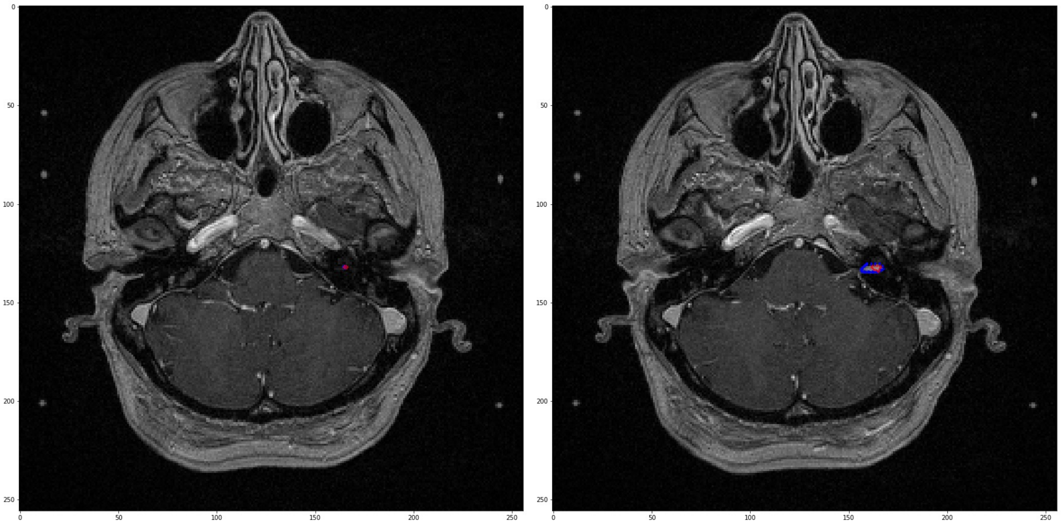
Figure 6: Top Prediction



(a) True volume: 5766.861 mm^3 Dice score: 0.682

(b) True volume: 56.5247 mm^3 Dice score: 0.731

Figure 7: Worst Prediction



(a) Slice: 59

(b) Slice:60

Figure 8: True volume: 253.2881 mm^3 Dice score: 0.747

4.4 Discussion

We have shown some examples that have high dice scores in Figure 5. The blue contour is the hand-drawn mask whereas the red one is our prediction. In Figure 5, the three cases that receive dice scores above 0.89 are the tumors either having a clear boundary with the other brain tissues or having greater contrasts to other areas. In more standard cases, tumors are usually much brighter thus making it easier for the model to detect.

There are a few tumors that our model does not perform well in masking their area. We find that when a tumor has hollow regions inside like in the left image of Figure 6, the model will only mask the highlighted place and ignore the dark area. Thus, in this case, the predicted mask is missing a part of the tumor thus resulting in a relatively low dice score. In addition, our model tends to perform less ideally when dealing with small tumors. The image is represented in a 3D matrix, pixel-wisely, and thus between slices, there will be thickness brought by the voxel size. Because dice loss is measuring how much the prediction overlaps the target, small tumors have a smaller volume and thus a little mismatch will introduce a rather large error. This thickness brought another issue presented in Figure 7. In the left image, the ground-truth mask only circles out a small dot for the tumor. However, on the next slice, the area that the ground-truth mask circles suddenly increase. This means that due to thickness between slices, we are missing information and the model fails to learn tumor distribution patterns from neighboring slices.

5 Conclusion

Currently, our model performs well with the single T1w sequence input. However, the original design of this U-net CNN model can take multiple MRI sequences by padding other tumor sequences as extra channels to the T1w sequence. For the current sequence, the tumor area is highlighted to increase contrasts to surrounding brain tissues, but T2 and CISS sequences darken the tumor area. As we can see from 4.3, one of the tumors that our model performs badly on is a heterogeneous tumor that has a hollow area inside, and thus our model could not identify the correct boundary. But when it comes to T2 and CISS sequences, the surrounding area will be lit up, and the whole tumor darkens out, which might be easier for the model to recognize when combined with the T1 sequence. However, T2 and CISS sequences are not available to us currently. Also, these sequences have different dimensions and voxel sizes compared to T1, which will require more complex pre-processing to be incorporated into the current model. Thus, incorporating T2 and CISS sequences of tumor masks can be future works to further optimize the model performance.

As we mentioned before, we developed an algorithm that calculates the tumor volume based on the 3D tumor mask. This algorithm will be crucial for future works when doing growth predictions. We want to classify each patient based on their tumor condition before and after the radiosurgery to determine treatment quality. One of the most important measurements will be checking whether the tumor grows or contracts.

6 Lessons Learned

The number one takeaway from this project is the importance of domain knowledge. As radiosurgery on brain tumors is an unacquainted field to all of our members, a large proportion of time is devoted to understanding the dataset and major background of data origins. With the domain knowledge acquired, we were able to perform later tasks efficiently and propose potential measures that meet specific requirements under our problem setting.

Another big takeaway is to prioritize data processing. Data sets for school work are mostly pre-processed and ready to use for modeling. However, in the industry, the fact turns over and the data processing part takes our team a large amount of time. In this case, we need to work with a special format, Nifty, for our brain scan images. We look into data for each patient and explore what features these images have. We also did a lot of search on related work, from which we learned what kind of processing a Nifty file may require. After discussing with our advisors, who are experienced in the field, we gained a full understanding of where we should pay attention to. Overall this was a long process as we are rotating from each stage of the data science parts.

7 Student Contribution

Implementation: Algorithm Development, Data Cleaning, Model Creation, Model Testing, Result Interpretation, Literature Review

Chengyu Chen, Bo Zhang, Xinyi Gu

Poster:

Xinyi: Background, Data Processing, Future Work

Bo: Model Architecture

Chengyu: Result, Conclusion

Report:

Xinyi: Introduction, Conclusion, Lesson Learned

Bo: Related Work, Problem Definition and Algorithm

Chengyu: Results Interpretation, Discussion

8 References

References

David G. Ellis and Michele R. Aizenberg. Trialing u-net training modifications for segmenting gliomas using open source deep learning framework. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 40–49, 03 2021. doi: 10.1007/978-3-030-72087-2_4.

Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus Maier-Hein. No new-net. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 234–244, 03 2019. doi: 10.1007/978-3-030-11726-9_21.

Bjoern H Menze, Andras Jakab, Stephan Bauer, and et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34:1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Guotai Wang, Wenqi Li, S ´ebastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. *Frontiers in Computational Neuroscience*, 13, 2019. doi: 10.3389/fncom.2019.00056.