Brain Tumor Segmentation from 3D MRI Scans via 2D U-Net

Annalisa Vitulli

Department of Computer Science

University of Central Florida

Annalisa.Vitulli@ucf.edu

1. Abstract

A 2D U-Net architecture is proposed for semantic segmentation via pixel-wise classification of multi-modal MRI images. Although any MRI modality could be used, slices from the T1Gd modality were used for training to maximize visualized tumor pixel exposure during model learning. K-fold cross-validation was used for training, and per fold, dice and Hausdorff distance results were reported along with the overall averages of the 5 folds.

2. Method

This 2D U-Net model for semantic segmentation of multi-modality MRI images is implemented to determine performance for brain tumor segmentation.

2.1. The 2D U-Net Architecture

Fully Convolutional Network. The implemented network follows the standard U-Net encoder-decoder architecture with skip connections, designed to capture spatial details and contextual information.

Encoder. Each encoder block consists of two convolutional layers, batch normalization, ReLU activation, and dropout ($dropout\_prob = 0.2$ per layer) [1].

Bottleneck. The bottleneck consists of a convolutional block with two convolutional layers, batch normalization, ReLU activation, and dropout ($dropout\_prob = 0.2$ per layer) [1].

Decoder. Each decoder block includes a convolution for up sampling, skip connections, followed by two convolutional layers, batch normalization, ReLU activation and dropout ($dropout\_prob = 0.2$ per layer) [1].

2.2. Data Preprocessing

Data Augmentation. Images were randomly cropped to between 80-120% of the original image's area and then resized to 128x128 pixels similar to previous methods for 2D U-Net implementations [1]. ColorJitter was applied with a brightness of .1, a contrast of .1, a saturation of 0.1, and a hue 0f 0.5 to simulate varying MRI modalities. RandomAffine (15° rotation, 10% translation of image's width and height, 90% to 110% scaling of image's original size, and 10° sheer. Random Rotation ($p = 0.5$), Random Horizontal Flip ($p = 0.5$), and Random Vertical Flip ($p = 0.5$) were applied.

Data Selection. Since the dataset consisted of 3D MRI scans, a single optimal slice was selected within the data loader based on the highest presence of tumor pixels [1]. This selected 2D slice was min-max normalized to scale the values between 0 and 1 to help improve model convergence.

2.3. Implementation Details

Dataset. A subset of data from the 2016 and 2017 Brain Tumour Image Segmentation (BraTS) challenges [2-4] was applied for the training and validation of this model. The BraTS dataset consists of pre-operative multimodal MRI scans with four imaging modalities (i.e., T1w, T1Gd, T2w, and FLAIR) of patients with gliomas−the most common type of brain tumor. In this implementation, 484 MRI scans−each scan containing the four formerly referenced modalities−were used for training and validation. The variation in MRI modalities is crucial for segmentation tasks since each modality provides varying details to the model's network. For example, T1-weighted images (i.e., T1w) aid in defining the anatomical structure of the brain while T2-weighted images (i.e., T2w) enhance edema and fluid [5]. The Fluid Attenuated Inversion Recovery (i.e., FLAIR) modality is especially beneficial at reducing noise caused by cerebrospinal fluid in the brain, allowing for improved delineation of the peritumoral edema and invasive tumor edges [5]. The T1 with Gadolinium (i.e., T1Gd) modality−sometimes called the T1 with contrast (i.e., T1C) modality−provides essential visualization of active tumor regions by enhancing blood-brain barrier disruption [5]. Figure 1 displays slices of the four different MRI modalities along with the corresponding segmentation map from a patient's MRI scan in the BraTS subset used for this implementation.

Training. The model was trained using k-fold cross-validation where $n\_splits = 5$.

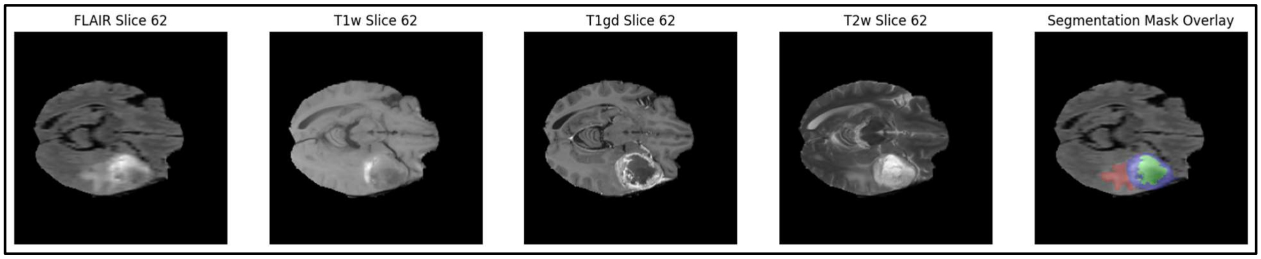

Figure 1. Multi-Modal Brain MRI Images with Tumor Segmentation Overlay

Slices from the T1Gd MRI modality were selected to train the model since previous studies in 2D U-Net implementations found the most success in identifying tumor regions with this modality [1]. Figure 2 and Figure 3 show MRI slice visualizations after data transformations were applied along with the corresponding annotated mask region.

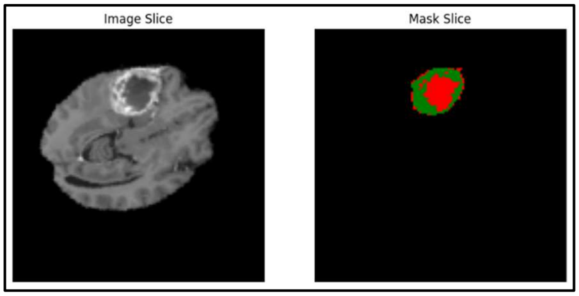

Figure 2. Brain MRI with Small Tumor Segmentation Mask

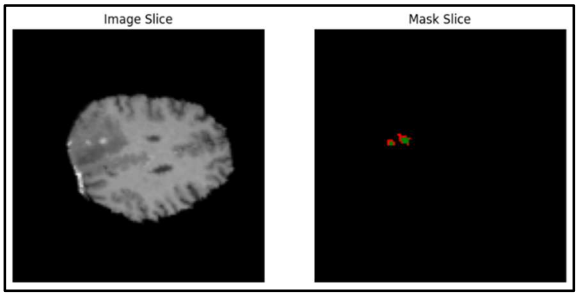

Figure 3. Brain MRI with Small Tumor Segmentation Mask

Hyperparameters. The following hyperparameters were used when training the model:

- **Epochs:** 50 [1].
- **Batch Size:** 16 [1].
- **Learning Rate (LR):** A LR of $1e^{-5}$ [1].
- **Regularization:** A weight decay of $1e^{-5}$.
- **Optimizer:** The AdamW optimizer.
- **Scheduler:** StepLR with a step size of 7 and gamma of 0.1 to reduce the LR periodically.
- **Loss Function:** Dice Loss with a smooth parameter of $smooth = 1$ to prevent division by zero and help stabilize training.

3. Results

Dice and Hausdorff distance segmentation results for the enhancing tumor regions (i.e., ET), regions of the tumor core (i.e., TC), and the whole tumor region (i.e., WT) were reported per fold along with the average results for each category over the 5 folds. The ET region consisted of Label 3 from the BraTS subset while the TC was comprised of Labels 2 and 3. The WT consisted of Labels 1, 2, and 3.

3.1. Metrics

Dice Score. The Dice scores for the ET region, TC region, and WT region were calculated using Equation 1 comparing the model's generated prediction to the provided ground truth segmentation masks.

$$Dice = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

Hausdorff Distance. The Hausdorff Distance for the previously mentioned labels was calculated as the greatest distance from a point in one set to the nearest point in the other set (e.g., the model's segmentation results to the ground truth). The medpy.metric implementation of Hausdroff Distance was used to calculate the value for this report [5].

The per fold segmentation results as well as the average of all folds' segmentation results are documented in Table 1.

FOLD	DICE_ET	HD_ET	DICE_TC	HD_TC	DICE_WT	HD_WT
1	0.000000	29.685269	0.000000	22.287458	0.195198	19.778444
2	0.000000	45.925753	0.048822	55.498627	0.326906	14.395896
3	0.116628	27.728204	0.082251	27.290511	0.308258	20.295563
4	0.000438	44.179817	0.000700	30.830586	0.150791	19.060384
5	0.000224	44.710141	0.170780	36.456596	0.210704	30.362822
OVERALL	0.023458	38.445837	0.060511	34.472756	0.238371	20.778622

Figure 1. Multi-Modal Brain MRI Images with Tumor Segmentation Overlay

3.2. Segmentation Accuracy

Dice Scores. Both the per fold and average Dice scores were incredibly low. As shown in Table 1, the average Enhancing Tumor, Tumor Core, and Whole Tumor Dice scores over the 5 folds were less than 1 when an acceptable result is above 50.

Hausdorff Distance. Similarly to the results observed in the Dice scores, the Hausdorff Distance demonstrated unfortunately poor model performance with scores ranging consistently in the 20s to low 50s across all classes (i.e., Enhancing Tumor, Tumor Core, and Whole Tumor).

Visualizations. Figure 4, Figure 5, and Figure 6 display MRI images along with the model's generated prediction followed by the radiologist labeled ground truth segmentation masks.

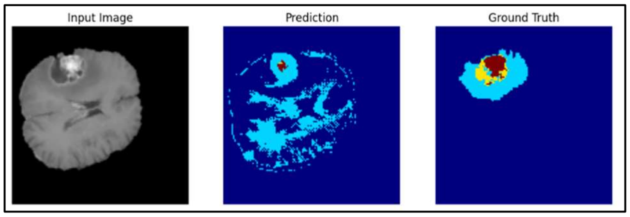

Figure 4. Misclassification with High Background Noise

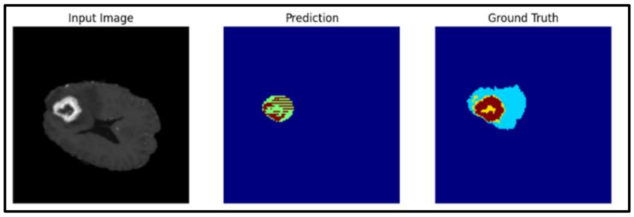

Figure 5. Misclassified Prediction of TC

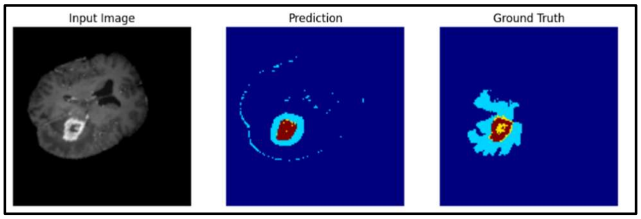

Figure 6. Semi Accurate TC Prediction with Minor Background Noise

4. Conclusions and Future Improvements

Although dropout was initially implemented to counteract overfitting, as shown in Figure 4 and Figure 5 as well as by the low Dice scores and high Hausdorff Distance values, the model struggled to learn any meaningful features from the training data. When analyzing Figure 5 for mispredictions, the model clearly failed to accurately distinguish between edema, non-enhancing, and enhancing tumor regions. Additionally, in Figure 4, although the model started to predict the tumor core in the correct location, the model inaccurately predicted edema across multiple locations.

Figure 6 demonstrates a slightly more successful prediction from the model although the misclassified edema is still prevalent. Since the model struggled to learn the tumor features accurately, moving forward, adding attention gates to the U-Net architecture to selectively refine feature maps could improve the performance. Also, Dice Loss could be combined with Categorical Cross-entropy Loss or Binary Cross-entropy loss for a hybrid loss approach. The latter two losses have both been tested in 2D U-Net implementations and shown successful results [1].

Lastly, instead of only providing a single MRI slice with the most non-zero label values to the model during training, multiple slices could be provided—anywhere between 20 to 30 which is roughly the quantity of slices a radiologist examines in brain MRIs when analyzing these tumor regions—in an attempt to provide the model with more information and to ideally improve learning.

References

[1] Bakas, S., Reyes, M., Int., E. & Menze, B. 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).

[2] Bakas, S. et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data 4, 1–13 (2017).

[3] Menze, B. H. et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 1993– 2024 (2015).

[4] Bonato, B.; Nanni, L.; Bertoldo, A. Advancing Precision: A Comprehensive Review of MRI Segmentation Datasets from BraTS Challenges (2012&ndash;2025). Sensors 2025, 25, 1838.

[5] Loli, "Hausdorff Distance Implementation in MedPy," MedPy Documentation, Accessed: Mar. 16, 2025. [Online]. Available: https://loli.github.io/medpy/_modules/medpy/metric/binary.html#hd