Journal of Integrated
SCIENCE & TECHNOLOGY

Estimation of uncertainty in Brain Tumor segmentation using modified multistage 3D-UNet on multimodal MRI images

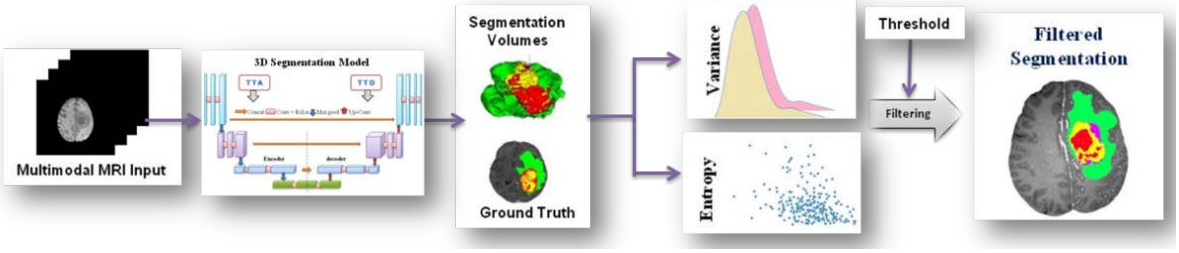
Bhavesh Parmar,[1*] Mehul Parikh,[2]

[1]*Gujarat Technological University, Ahmedabad, India.* [2]*L.D. College of Engineering, Ahmedabad, Gujarat, India.*

Received on: 27-Nov-2023, Accepted and Published on: 02-Feb-2024

Article

ABSTRACT



Automated brain tumor segmentation is challenging due to the tumor tissues' shape, size, and appearance. Various methods used multi-mode MRI scans to segment sub-regions of brain tumors. 3D CNN methods improved performance in recent years, but most methods do not use uncertainty information in segmentation. For reliability and understanding, model prediction is vital for clinical decisions. This work studies three models namely 3D-UNet, Modified 3D-UNet, and Modified Multistage-3D-UNet for brain tumor segmentation. MRI volume bias correction and normalization were carried out using z-score normalization. Two patch generation strategies reduce memory use and class imbalance. Voxel-wise uncertainty evaluation was made for aleatoric and epistemic uncertainties using test time augmentation and dropout, respectively. Variance and entropy are used to measure the uncertainty of the modified multistage-3D-Unet segmentation model from ground truth. Variance creates separate uncertainty maps for each tumor sub-regions, whereas, entropy provides only global information. Uncertainty is used to filter miss-segmented predictions and improve accuracy. Uncertainty awareness increases model accuracy with dice scores of 0.93, 0.91, and 0.83 for tumor sub-regions WT, TC, and ET respectively.

Keywords: brain tumor segmentation, uncertainty, Magnetic Resonance Imaging, 3D UNet, Deep Learning

INTRODUCTION

A brain tumor is an abnormal development of tissue in central nervous system. The tumor can be benign or malignant. Development of tumor in brain known as primary brain tumor and spread of tumor to brain tissue from other parts of body is metastatic or secondary brain tumor.[1] Brain tumors can be classified based on the tissue type, anatomical location and malignancy level. World Health Organization (WHO) classifies brain tumor based on histological features, molecular characteristics and malignancy. The grading ranges from Grade I to Grade IV with increasing malignancy or severity.[2] Gliomas are the most common type primary brain tumor.[1] Glioma can be classified into two types (1) “low grade gliomas" (LGG), grades I-II, characterized by low blood concentration with less aggression, and (2) “high grade gliomas" (HGG), grades III-IV, with faster growth rate and aggressiveness.[2]

Different MRI modalities (T1-weighted, post-contrast T1-weighted, T2-weighted, and Fluid-Attenuated Inversion Recovery-FLAIR) reflect the heterogeneous tumor sub-region properties for diagnosis, treatment and evaluation of brain tumor.[3] The diverse brain tumor sub-regions are peritumoral edematous / invaded tissue, the fluid filled necrotic core, the enhancing solid non-enhancing tumor core. Manual segmentation of MRI modalities is challenging and time consuming for clinical experts. There is a need of automated method for automated brain tumor segmentation. Despite the advancement in recent years, due to highly diverse size, appearance, and shape of malignancy and reliability of available methods, Developing trustworthy automated multi-modal brain tumor segmentation is a challanging task.[4]

The work presented here is regarding the development of algorithm for finding uncertainy in brain tumor segmentation using 3D multi-modal Magnetic Resonance Imaging (MRI) scans. The proposed method is implemented with well accepted 3D-UNet[5]

*Corresponding Author: Bhavesh Parmar, Research Scholar, Gujarat Technological University, Ahmedabad, and Dr. Mehul Parikh, L. D. College of Engineering, Ahmedabad, India – 380 015
Email: bhparamar@gmail.com, mehulcparikh@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(5), 802.*
URN:NBN:sciencein.jist.2024.v12.802

Modified 3D-UNet[6] and modified 3D-Unet with multistage training. Two different patch generation techniques proposed to decrease the class imbalance problem. Patch based sampling is used due to memory restriction and data augmentation carried out to avoid model overfitting.

Aleatoric[7] and epistemic uncertainties have been estimated from the segmentation predictions using Test Time Augmentation (TTA) and Test Time Dropout (TTD) respectively. Voxel wise uncertainty measure in terms of variance and entropy gives tumor sub-region and global information in segmentation prediction to that of ground truth. Hybrid uncertainty derived from the combination of aleatoric and epistemic uncertainties on modified Multistage-3D-UNet. False positive predictions were filtered out using suitable threshold to improve model performance in segmentation prediction in different classes.

RELATED WORKS

In the last decade many researchers have focused on medical image analysis problems, particularly developing methods for efficient brain tumor segmentation. Existing brain tumor segmentation approaches can be generalized in two classes; namely generative and discriminative. Generative methods use preceding acquaintance and probabilistic distribution of model. Discriminative methods extract features from dataset. With advancement in Convolutional Neural Network (CNN), later appraoches have increase in recent algorithms of brain tumor segmentation. DeepMedic[5,8] and nnU-Net[9] were the breakthrough in the area of medical imaging using CNN.

For clinical decision making and to help understand the reliability of segmentation prediction, uncertainty information is important. Many studies (Eaton-Rosen et al., 2018; Jungo et al., 2018; Wang et al., 2019a; Wang et al., 2019b) exploit medical uncertainty based aleatoric or epistemic estimations.[10-13] Wang et al. proposed a cascade of hierarchical CNNs to segment all brain-tumor structures, and used test-time augmentation to obtain not only segmentation outputs but also data-based uncertainty (aleatoric) of all structures of brain-tumor segmentation.[13] In another study Wang et al. used a combination of aleatoric and epistemic to estimate uncertainties for whole tumor segmentation.[12] Finally, McKinley et al. proposes to incorporate uncertainty measures during training by defining loss function that models noise and uncertainty.[14]

There is a need of method, which can address measurement of segmentation uncertainty in all brain sub-regions along with global information for aleatoric and epistemic uncertainty measurement and provide segmentation with more confidence.

METHOD

Dataset Description

This work use BraTS2020 dataset of Brain Tumor Segmentation (BraTS) Challenge 2020.[15] The multimodal BraTS2020 dataset provides four different scans, namely T1 (native), T1ce – (post-contrast T1-weighted), T2 (T2 weighted), and FLAIR (T2 Fluid Attenuated Inversion Recovery). The data acquired routine clinically acquired multimodal MRI from multiple institutions by means of non-identical clinical protocols and different scanners.[16-18] The multimodal MRI scans containing gliomas with segmentation tumor as ground truth annotations by expert neuroradiologists, with pathological resolute opinion. The annotations distinct by the non-enhancing tumor core and necrotic (NET/NCR - label 1), the peritumoral edema (ED - label 2), and the GD-enhancing tumor (ET - label 4).[16,19,20] BraTS dataset includes 369 cases (293 HGG and 76 LGG). The 3D MRI of all four modalities are rigidly aligned, resample to 1 mm^3 isotropic resolution and skull-stripped. Each modality MRI volume has size of *240 x 240 x 155*. Validation and test sets include 125 and 166 cases respectively.

The comparative small region of brain tumor to whole brain tissue in MRI creates class imbalance. In the Brain volumes, tumor region accounts 5-15 % with further smaller sub-tumor classes.[21] The volume distribution of each tumor sub regions class presented in figure 1. In global sub region distribution evidently tumor sub class ED is more likely to occur, compare to other two classes, ET and NCR. The NCR shows high variability between subjects. Global sub region distribution is aligned with distribution in HGG glioma subjects. LGG subjects having appaerently low ET region voxels compare to other two sub region ED and NCR distribution. The low ET appearance can be justified by the LGG tumors with low blood concentration.

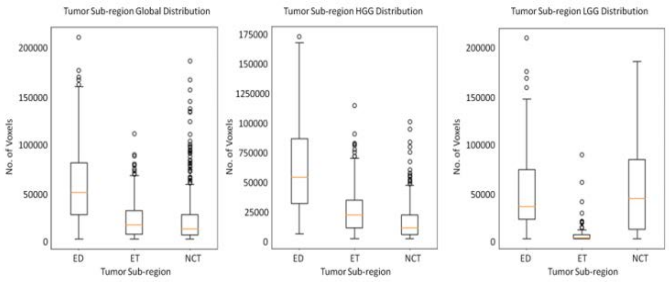


Figure 1. Voxel distribution in ED, ET, NCR with class imbalance.

Data Preparation

The MRI scans having inherent bias, due to variation in magnetic field with multiple scanner, clinical protocols, and multi-institutional nature of data. The structural MRI bias field correction was carried out with N4ITK Insight Toolkit.[22] Each MRI modality volumes were normalized and standardized with Z-score normalization and min-max scaling. Mean value and the standard deviation are estimated on all training volumes by accumulating the voxel intensity values of the brain. The final value of each voxel pertaining to brain would range from 10 to 110. Background voxels distinguished with value 0. Similar set of values of mean and standard deviation were used in validation and testing phase in Z-score normalization process. Each MRI sequences are normalized in 0 to 1 range of voxel value before feeding for training. To surmount the problem of overfitting, augmentation was implemented with least data disruption. Data augmentation was done with Random intensity shift between (-0.1, 0.1), random rotation on two axis with 50% probability, and random flip on all three axis with 50% probability. During training phase, random axis mirror along the horizontal axis was also applied for augmentation.

Sampling Strategy

Larger MRI volumes cannot be feed directly to 3D-CNNs due to its computational complexity. To increase the images fed in single

batch, patch-base strategis are effective.[6] Bigger batch size could increase optimization but leads less contextual information with smaller patches. Whereas, bigger patches give more contextual information, leads to the smaller batch size, which increases the variance of stochastic gradient and reduces optimization. To overcome the trade-off and taking advantage between batch and patch size, training batch pool with various patch size was created as preset. Different padding and cropping layers between convolution layers in network can learn global information from the bigger patch and informative texture from small patch with the same parameter.

Two different patch generating strategy were used to deal with larger MRI volumes on a less powerful GPU. Two different strategies based on cuboidal boundary were considered for generation of patches of size 128 X 128 X 128.

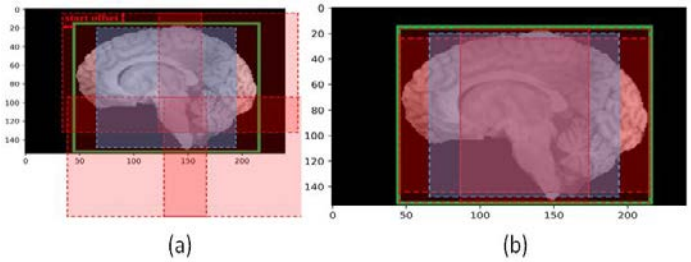


Figure 2. Two strategies for cubical patch generation.

The first strategy as shown in figure 2(a), generate cubic patch with origin at a random starting point on boundary at distance ranges 0 - 6 voxels away. Keeping 32 voxel overlap between neighboring patches, next patch will be generated by moving patch window with 96 voxel from the origin of previous patch. This strategy will also produce the background values. The second strategy as shown in figure 2(b) created patch starting at the corner of the quadrilateral boundary of scan. Next patches will be generated in the same manner to former method. The second strategy arrange all patches to a wide extent inside the brain. Patching stretegies are also usefull in maintaining the inter class distribution.

Loss Function

Earlier methods suggest the weighted multi-class dice loss function for deep learning based brain tumor segmentation in earlier BraTS.[17] This MRI dataset has a significant class imbalance. Voxels representing sub-regions of brain tumors are outnumbered by those containing healthy tissues. Outliers are not taken into account after the dataset is cleaned. To address class-imbalance data, the customized weighted multiclass dice loss function can be utilized during network training, as proposed in Bakas et al.[17] and other existing methods. The inherent class imbalance problem was addressed using weighted multi-class dice loss function in the network.

$$L_{Dice} = -\sum_{c=0}^{2} \frac{\sum_{i,j,k=0}^{127} G_{c_{ijk}} Y_{c_{ijk}}}{\sum_{i,j,k=0}^{127} G_{c_{ijk}} + \sum_{i,j,k=0}^{127} Y_{c_{ijk}}} \quad (1)$$

Where c is channel number, $c \in \{0;2\}$. Ground truth image matrix is G of size *3 X 128 X 128 X 128*. Y is the output of the network and i, j, k are voxel locations with values $\{i, j, k\} \in \{0;127\}$.

Network Architecture

This work implements two networks, 3D U-Net[5] and Modified 3D U-Net[6] architectures, for brain tumor segmentation and creates an ensemble to reduce the bias in each independent model. The modified 3D U-Net[6] is the improvement over 3D U-Net.[5]

3D-UNet The original implementation with some minor modifications is implemented. Batch Normalization is changed for Group Normalization with 32 feature maps at the highest resolution.

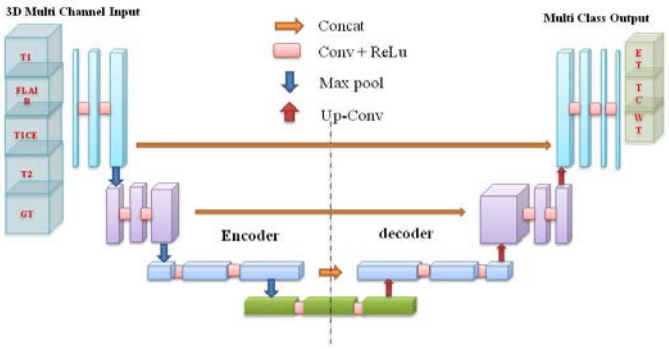


Figure 3. 3D-Unet architecture with encoder-decoder blocks [5].

The network architecture is divided into symmetric Encoder and Decoder parts. The Encoder is includes two convolutional blocks consisting of 3DConv + ReLu + GroupNorm structure. The downsampling and upsampling is performed with 2^3 Max-Pooling and interpolation respectively. Convolutional layers have kernel size *3 x 3 x 3*, and the last convolutional layer reduced to *1 x 1 x 1* kernel size with 4 feature maps for multi-class output in terms of segmentation labels. ReLu non-linearity and the skip-connections are joined with a concatenation step. The network outputs a four-channel segmentation map with the training labels as well as a softmax. The detailed architecture can be seen in Figure 3.

Modified 3D-UNet The basic 3D U-Net[5] expanded with residual connections. This will be decresing the vanishing gradient problem by allowing a deeper network. Apart from the residual blocks, other modifications compare to basic 3D U-Net, in network structure includes, (i) upsampling changed to interpolation for transposed convolution, (ii) element-wise addition to join skip-connections.

As shown in figure 3 and 4, in both the network, all three modalities along with ground truth are input to the network as *4 x 128 x 128 x 128* matrices stack with patch size *128 x 128 x 128*.

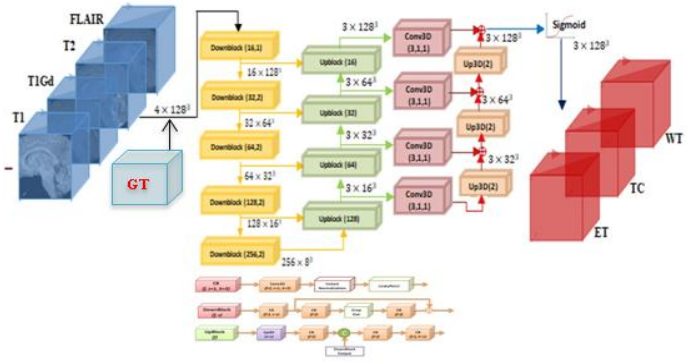


Figure 4. Modified 3D-Unet architecture [6].

Downblocks reduce patch size and increase the channel length, whereas, convolution block (CB) follows instant normalization and leaky ReLU.[23] Downblock pass information from front to end to eradicate overfitting,. The upblock reconstructs the location information by joining corresponding downblock outputs. The patch size and number of channels are recovered with a probability matrix with the confidence of each voxel belongs to a particular sub-region of the tumor, namely WT, TC, and ET.

This study made use of three distinct 3D UNet segmentation architectures. Original 3D UNet suggested by Isensee et al. Modified 3D UNet by Parmar et al. and proposed modified 3D UNet with multistage training. As input, the first two techniques took a 112 x 112 x 112 patch size and a batch size of 2. In contrast, the multistage-modified 3D UNet network underwent a two-stage training process. At first, with a 128 × 128 x 128 patch size for 300 epochs and batch size of 1, and then, for 200 epochs in batch size of 2, with a 112 x 112 x 112 patch size. Modified 3D UNet was trained with two-stage training (Multistage-3D-UNet), considering the sampling strategy. As a result, the network may be trained more effectively using both local and global data.

The all three networks were trained using the ADAM optimizer, with initial learning rate of 1×10^{-4}. For unchanged validation loss over last 30 epochs, learning rate will decreased by a factor of 5. Training was regularized weight decay of 1×10^{-5}.

A single patch is created by combining all output channels of patch size *128 x 128 x 128*. Incorporate with suitable threshold values, the intensities of each voxel in the patch represents one of the tumor sub-region or background.

Post-processing

Post-processing step was carried out to reduce the false positive, small and separated components. Suitable threshold value was set prom the training set knowledge to correct the false positive. Connected components with larger proportion than threshold values were kept. This process will leads to removal of false positive smaller and separated components but larger components can be kept, in view of the fact that some subjects may have multiple separate tumors.

Most of the existing methods expressed the biggest intricacy of providing accurate segmentation of the smallest sub-region ET. ET sub-region is hard to segment in LGG subjects. Almost 40% subjects missing enhancing tumor.[16] As suggested in nnNet,[5,9] ET volume lower to the threshold can be replaced to necrosis to improve the overall accuracy of model. Such threshold values can be set through independent experiment.

Uncertinty

The work carried out to model voxel-wise uncertainty using Test Time Augmentation (TTA) and Test Time Dropout (TTD) for aleatoric and epistemic uncertainties respectively.

Aleatoric uncertainty was model in similar manner using augmentation techniques in training steps with adding up random Gaussian noise. This process can add modifications not previously encountered by the network.

Epistemic uncertainty was computed as proposed by Gal et.al.[24] For task simplification Bayesian Approximation was used fro dropout, during both, training and tastig phase. Further, it is suggested to repeat the prediction several hundred times with random dropout. The final prediction can be averaged out over all predictions. For $Y^i = \{y_1^i, y_2^i \dots y_n^i\}$, vector representing ith voxel prediction label. For each evaluation regions, the uncertainty can be modeled by finding the variance $(\sigma^2)$of each prediction as expressed in equation 2.

$$\sigma^2 = \frac{1}{N}\sum_{n=0}^{N}\left(y_n^i - y_\mu^i\right)^2 \qquad (2)$$

where μ indicates mean value of i^{th} voxel predections with N iterations. 20 iterations with 50% random dropout probability used to eliminate channel. For each sub-region, the uncertainty maps were generated independently.

As suggested by Wang et.al.,[13] Uncertainty can also be predicted using entropy parameter. The voxel-wise uncertainty using entropy (*S*) can be calculated as:

$$S\left(Y^i|X\right) \approx -\sum_{m=0}^{M}\hat{p}_m^i \ln\left(\hat{p}_m^i\right) \qquad (3)$$

where $\hat{p}_m^i$ is the frequency of the m^{th} unique value in Y^i and X represent the input image. Global uncertainty map can be generated using entropy.

Experimental Setup

Uncertainty estimations were carried out considering three strategies to evaluate the model behaviour in different conditions. In first experiment, aleatoric uncertainty was modeled with test time augmentation (TTA). The second experiment was designed to model epistemic uncertainty with test time dropuout (TTD). Whereas, the third experiment model hybrid uncertainyy (aleatoric + epistemic) with TTA and TTD together. For all three experiments variance and entropy parameters were used to predict the uncertainty maps. The final prediction and uncertainty maps are computed with same strategies for aleatoric and epistemic uncertainties. Entropy will provide a single global measure of uncertainty, while, each sub-region uncertainty maps were generated using variance.

RESULTS

The algorithm has been implemented in python using Pytorch[25] and TensorFlow[26] on NVIDIA Quadro P5000 GPU for training, validation and testing of model. All the previously mentioned Convolutional Neural Network (CNN) architectures were initially trained on 295 (80 %) data (divided in training (280 images) and validation (15 images) set) and tested on 74 (20 %) data. The models were further trained on 369 scans, validated on 125 scans and tested with 166 previously unknown images using BraTS dataset. Dice score, Hausdorff distance (95^{th} percentile), sensitivity and specificity for each class were evaluated. Ratio of filtered TN (FTN) and ratio of filtered TP (FTP) are used for specific uncertainty evaluation metrics.

Segmentation

Sensitivity (the rate of true positives) and specificity (the rate of true negatives) for the voxels were determined. The degree to which the segmented areas overlap voxel-wise was measured by sensitivity, specificity, and dice score. The surface distance, or the distance between segmentation borders, was assessed by a separate set of scores.The Dice coefficient can be used to determine how well a predicted segmentation matches up with the ground truth

voxel-wise. Multiplying the intersection area by the total number of voxels in both regions yields the dice coefficient.

Evaluation of segmentation performance can be measured with also known as the Sørensen–Dice index also known as the Dice Similarity Coefficient (DSC) of simply Dice.[27] DSC represents the degree of overlap between predicted region map and ground truth and calculated using equation 4 as;

$$DSC = \frac{2\,|P \cap Q|}{|P| + |Q|} \quad (4)$$

Where P represents predicted values and Q is the corresponding ground truth. The symbol | • | denotes the volume of underlying region. Considering Boolean nature of DSC for predicted label and ground truth at voxels level, sub-region dice can be calculated using equation 5 as;

$$DSC = \frac{2TP}{FP + 2TP + FN} \quad (5)$$

Where *TP*, *FP* and *FN* are values of True Positive (TP), False Positive (FP) and False Negative (FN) respectively. Given that the problem involves more than one class, equation (5) is a modified Dice equation (4). In this context, the concepts of sensitivity and specificity are examined via true positive and true negative values.

Hausdorf distance measures the maximum distance of one set to the nearest point in the other set [18], defined as:

$$D_H(P,Q) = max\left\{\left(sup_{x\epsilon P} inf_{y\epsilon Q} d(x,y)\right), \left(sup_{y\epsilon P} inf_{x\epsilon P} d(x,y)\right)\right\} \quad (6)$$

Where *sup* and *inf* represents the supremum and the infimum among the considered sets. Hausdorff distance at 95th percentile (HD95) is considered to avoid noisy segmentation and achieve more robust results the evaluation scheme uses the 95th percentile.

Table 1. Segmentation Results on Training phase (369 cases).

Method	Dice			Hausdorff (mm)		
	WT	TC	ET	WT	TC	ET
3D-UNet	0.85	0.84	0.76	6.97	10.13	28.23
Modified 3D-UNet	0.82	0.82	0.76	8.56	12.11	28.93
Modified Multistage-3D-UNet	0.87	0.86	0.79	9.19	11.89	30.94
Ensemble - mean	0.87	0.85	0.79	9.46	11.90	29.03

Segmentation results for training phase of 369 training data are presented in table 1 using Dice and Hausdorff Distance (95th Percentile) for 3D UNet, modified 3D UNet, modified Multistage-3D-UNet, and ensemble mean of all considered network

Table 2. Segmentation Results on Validation phase (125 cases).

Method	Dice			Hausdorff (mm)		
	WT	TC	ET	WT	TC	ET
3D-UNet	0.84	0.83	0.76	7.37	11.15	26.78
Modified 3D-UNet	0.83	0.82	0.74	5.34	11.34	23.29
Modified Multistage-3D-UNet	0.88	0.85	0.76	8.43	11.37	23.09
Ensemble - mean	0.88	0.85	0.77	8.43	11.90	21.04

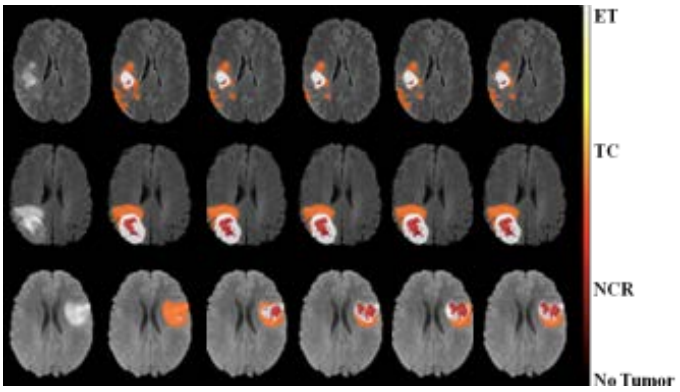
Table 2 represent the evaluation parameters during validation phase with 125 data.

The evaluation of segmentation model was carried out on all trained network models to understand the model behaviors. of test sets was carried out. Table 3 represents the output of modified Multistage-3D-UNet architecture with post processing in training, validation and test phases on 166 cases for comparasion purpose.

Table 3. Segmentation Results modified multistage-3D-UNet with post-processing.

Dataset	Dice			Hausdorff (mm)		
	WT	TC	ET	WT	TC	ET
Train	0.87	0.86	0.79	9.19	11.89	30.94
Valid	0.88	0.85	0.76	8.43	11.37	23.09
Test	0.89	0.86	0.80	8.26	11.11	23.83

In the absence of ET label in ground truth, all models were highly penalized. It was observed that, 3D-UNet based model gives false negatives. Use of small patches in place of whole tumor volume causes false positives. This aggravates variation in the ratio of healthy tissue against tumor classes. Contrary, larger number of false negatives may be present due to use of bigger patch sizes and pooling layers instead of strid convolutions in 3D-UNet models.



(a) (b) (c) (d) (e) (f)

Figure 5. Training results on patients: 115, 175, and 310 (top-bottom). For (a) Flair (b) Ground Truth (c) Modified Multistage-3D-UNet (d) Modified 3D-UNet (e) 3D-UNet (f) Ensemble mean.

Increasing the patch size may leads to lessen false positives but it fail to see local information, which cause label miss-classification on the region's boundaries. Qualitative segmentation results of various sample subjects are shown in figure 5 for visual comparison of models.

Uncertainty

Three uncertainty maps along with corresponding prediction maps were generated, one for each sub-region (WT, TC, ET). Normalization was applied on voxel values between 0-100, indicating most certain prediction with “0” and “100” represent most uncertain predictions. To calculate uncertainty in each class, Filtered True Positive (FTP) ratio and Filtered True Negative (FTN) ratio where used. FTP ration is defined as;

$$FTP = \frac{(TP_{100}/TP_T)}{TP_{100}} \quad (7)$$

Where more uncertain values are filtered using threshold *T*. The FTN ratio can be also calculated in a similar manner. The integrated score can be calculated as follows:

$$score = AUC_1 + (1 - AUC_2) + (1 - AUC_3) \qquad (8)$$

In the above equation 8, AUC_1, AUC_2, *and* AUC_3 are area under curves for dice, FTP and FTN parameters respectively.

Modified Multistage-3D-UNet outperformed to other considered networks with balanced results among various sub-regions. Table 4 present the results with aleatoric, epistomic and hybrid uncertainty considerations computed with variance and entropy parameters. The Dice value calculated here is computed by averaging the segmentation output for several thresholds used for filter uncertain predictions, It is evident from the displayed results that, Dice obtained with filtered uncertainty prediction improves to that of earlier segmentation on each sub-regions (WT : 0:92, TC : 0:89, ET : 0:82). Further, the model was more certain on the TP and less certain on FP and FN. AUC-Dice gives higher values while using entropy as the uncertainty measure.

Table 4. Results with estimate uncertainty on modified Multistage-3D-UNet during training phase.

Measure	Method	Dice			Ratio FTP			Ratio FTN		
		WT	TC	ET	WT	TC	ET	WT	TC	ET
Variance	TTA	0.91	0.89	0.82	0.02	0.04	0.03	8.0e-4	2.0e-4	1.8e-4
	TTD	0.90	0.87	0.80	0.17	0.15	0.08	2.3e-3	1.6e-3	3.0e-4
	Hybrid	0.92	0.88	0.82	0.17	0.15	0.09	2.6e-3	2.1e-3	4.0e-4
Entropy	TTA	0.92	0.88	0.81	0.05	0.04	0.05	1.2e-3	4.6e-3	5.3e-3
	TTD	0.91	0.87	0.81	0.14	0.12	0.07	2.2e-3	7.2e-3	1.2e-2
	Hybrid	0.92	0.89	0.82	0.16	0.11	0.06	3.1e-3	1.1e-3	1.9e-2

Results show that the model is more uncertain on epistemic uncertainty in LGG patients. This can be solved by introducing more training data to achieve more confident predictions. From the results and comparing different uncertainties, its obvious that; Aleatoric uncertainty focus on the region boundaries, with small variations, epistemic improves results on the ET region but filters more TP and TN, and the hybrid approach accomplish the best Dice-AUC results when entropy is used as the uncertainty measurement.

Table 5 shows the results in Training, validation, and test sets. The achieved integrated scores for validation are 0.91, 0.87 and 0.81 and for test 0.93, 0.91, 0.83 for WT, TC and ET respectively. The improvement of up to 2 point can be observed on the ET and TC sub-regions for test set.

Table 5. Uncertainty Results for the modified Multistage-3D-UNet.

Dataset	Dice			Ratio FTP AUC			Ratio FTN AUC		
	WT	TC	ET	WT	TC	ET	WT	TC	ET
Train	0.9212	0.8915	0.8178	0.0349	0.0238	0.0580	0.0008	0.0002	0.0001
Valid	0.9116	0.8715	0.8088	0.0422	0.0333	0.0380	0.0009	0.0002	0.0001
Test	0.9299	0.9084	0.8254	0.0322	0.0536	0.0375	0.0019	0.0004	0.0003

DISCUSSION

Cancer is one of major ailments that are responsible for a number of deaths every year.[28] The early diagnosis of the cancer is the key factor in controlling of this ailment.[29] MRI and other diagnostic techniques are at the fore-front for diagnosis while the proper detection of cancer from images still remains a challenge.[30] In this work three set of models on variations of 3D-UNet based CNNs specialized in medical imaging are proposed. 3D-UNet, Modified 3D-UNet, and Modified Multistage-3D-UNet outperforms in particular tumor sub-regions. Ensemble of these models was also defined to increase the model performance in segmentation. To increase the trustworthiness, consistency and understanding of the model, uncertainty implications on predicted segmentations were analyzed. Uncertainty prediction was also used to consider with suitable threshold to filter out prediction with less certainty, in turn increase the model accuracy in segmentation. Uncertainty measures were implemented on modified Multistage-3D-UNet.

The best outcomes of segmentation were achieved in ensemble of all three models, as the limitations of each model can be controlled. The results are comparatively good with the existing methods, still improvement chance can be sought. Presented results can still be modified with higher accuracy by changing training strategies for correct label distribution, thereby reducing false detections. The absence of ET regions in ground truth creates uncertain predictions in all models, which gratly decrease the accuracy. In order to improve the results, future work will be carried out for better representation of labels to the network with better viewability of local and global information in dataset than current patch strategies.

The results of all three segmentation models achieve good results. By adding more complexity to the network at different layers may boost the model performance. It can be also observed that LGG patient MRI shows low accuracy due to blood perfusion in surrounding tissues. Post-processing methods and targeted training strategies for each glioma grades can be helpful in differentiating sub-region segmentation predictions.

In this work, TTD use random dropouts, which can be structured with improved with Monte Carlo Dropout or any other well defined method to prevent loosing relevant label information. Also there is a scope of improvement in augmentation techniques. Variance and entropy are considered for uncertainty estimation. Other statistical parameters like Z-score can be explore to find the uncertainty prediction.

CONCLUSIONS

This work focus on improvement of segmentation of tumor sub-regions considering the uncertainty assessment. 3D-UNet, modified 3D-UNet and modified Multistage-3D-Unet (modified 3D-UNet with 2 stage training) models were trained on BraTS2020 datasets. Aleatoric, epistemic and hybrid uncertainty in the model segmentation prediction were evaluated. Uncertainty estimation carried out using variance, to evaluate uncertainty in each tumor sub-regions and entropy, as global measure. Uncertainty predictions were used to filter false positive segmentation with empirical threshold values for different tumor sub-regions. The result presented here, shows that uncertainty awareness leads to improvement in segmentation accuracy. The best brain tumor

segmentation results were achieved with Dice score with hybrid approach using entropy as uncertainty prediction. The results achieved with uncertainty aware segmentation in this methods for WT, TC and ET as Dice score of 0.93, 0.91, and 0.83 respectively in test set. The results can be further improved by implementing different network training strategies and using structured dropout and augmentation strategies during test time.

CONFLICT OF INTEREST STATEMENT

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper..

REFERENCES

1. F. Khodamoradi, M. Ghoncheh, R. Pakzad, H. Gandomani, H. Salehiniya. the Incidence and Mortality of Brain and Central Nervous System Cancer and Their Relationship With Human Development Index in the World. *World Cancer Res. J.* **2017**, 4 (4), 5.
2. C. Villa, C. Miquel, D. Mosses, M. Bernier, A.L. Di Stefano. The 2016 World Health Organization classification of tumours of the central nervous system. *Press. Medicale* **2018**, 47 (11-12P2), e187–e200.
3. J.R. Fink, M. Muzi, M. Peck, K.A. Krohn. Multimodality Brain Tumor Imaging: MR Imaging, PET, and PET/MR Imaging. *J. Nucl. Med.* **2015**, 56 (10), 1554–1561.
4. M.R. Patel, V. Tse. Diagnosis and staging of brain tumors. *Seminars in Roentgenology*. 2004, pp 397–411.
5. R.R. Agravat, M.S. Raval. Brain tumor segmentation and survival prediction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Quebec, Canada, **2020**; Vol. 11992 LNCS, pp 338–348.
6. B. Parmar, M. Parikh. Brain Tumor Segmentation and Survival Prediction Using Patch Based Modified 3D U-Net. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Crimi, A., Bakas, S., Eds.; Springer, Cham, **2021**; Vol. 12659 LNCS, pp 398–409.
7. A. Der Kiureghian, O. Ditlevsen. Aleatory or epistemic? Does it matter? *Struct. Saf.* **2009**, 31 (2), 105–112.
8. K. Kamnitsas, C. Ledig, V.F.J. Newcombe, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, 36, 61–78.
9. F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K.H. Maier-Hein. No New-Net; International MICCAI Brainlesion Workshop, Springer, **2019**; pp 234–244.
10. Z. Eaton-Rosen, F. Bragman, S. Bisdas, S. Ourselin, M.J. Cardoso. Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2018**, 11070 LNCS, 691–699.
11. A. Jungo, R. Meier, E. Ermis, E. Herrmann, M. Reyes. Uncertainty-driven Sanity Check: Application to Postoperative Brain Tumor Cavity Segmentation. In *1st Conference on Medical Imaging with Deep Learning*; **2018**.
12. G. Wang, W. Li, M. Aertsen, et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **2019**, 338, 34–45.
13. G. Wang, W. Li, S. Ourselin, T. Vercauteren. Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation. *Front. Comput. Neurosci.* **2019**, 13 (56), 056.
14. R. McKinley, R. Meier, R. Wiest. Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2019**, 11384 LNCS, 456–465.
15. S. Bakas, B. Menze, C. Davatzikos, et al. MICCAI BRATS - The Multimodal Brain Tumor Segmentation Challenge 2020. 2020.
16. B.H. Menze, A. Jakab, S. Bauer, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, 34 (10), 1993–2024.
17. S. Bakas, M. Reyes, A. Jakab, et al. Identifying The Best Machine Learning Algorithms for Brain Tumor Segmentation. In *arXiv preprint arXiv:1811.02629*; **2018**.
18. Y.-X. Zhao, Y.-M. Zhang, C.-L. Liu. Bag of Tricks for 3D MRI Brain Tumor Segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer, Cham, **2020**; Vol. 11992 LNCS, pp 210–220.
19. P. Deshmukh, V.R. Pawar, A.N. Gaikwad. Machine learning based approach for lesion segmentation and severity level classification of diabetic retinopathy. *J. Integr. Sci. Technol.* **2023**, 11 (4), 576.
20. B. Mehta, V. Dave. Detection of Cardiovascular Autonomic Neuropathy using Machine Learning Algorithms. *J. Integr. Sci. Technol.* **2023**, 11 (4), 562.
21. L.M. Ballestar, V. Vilaplana. MRI Brain Tumor Segmentation and Uncertainty Estimation Using 3D-UNet Architectures. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer, Cham, **2021**; Vol. 12658 LNCS, pp 376–390.
22. N.J. Tustison, B.B. Avants, P.A. Cook, et al. N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* **2010**, 29 (6), 1310–1320.
23. Z. Fan, Y. Jiang, P. Wang, et al. Unified Implicit Neural Stylization; **2022**; pp 636–654.
24. Y. Gal, Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd International Conference on Machine Learning, ICML 2016*. 2016, pp 1651–1660.
25. A. Paszke, S. Gross, S. Chintala, et al. Automatic differentiation in pytorch; NIPS-W, **2017**.
26. D. Jha, P.H. Smedsrud, M.A. Riegler, et al. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In *Proceedings - 2019 IEEE International Symposium on Multimedia, ISM 2019*; San Diego, CA, USA, **2019**; pp 225–230.
27. A.D. Yao, D.L. Cheng, I. Pan, F. Kitamura. Deep Learning in Neuroradiology: A Systematic Review of Current Algorithms and Approaches for the New Wave of Imaging Technology. *Radiol. Artif. Intell.* **2020**, 2 (2), e190026.
28. B.S. Chhikara, K. Parang. Global Cancer Statistics 2022: the trends projection analysis. *Chem. Biol. Lett.* **2023**, 10 (1), 451.
29. M. Kumari, B.S. Chhikara, P. Singh, B. Rathi, G. Singh. Signaling and molecular pathways implicated in oral cancer: A concise review. *Chem. Biol. Lett.* **2024**, 11 (1), 652.
30. B. Kalva, M.C. Mohan. Effective image retrieval based on an optimized algorithm utilizing a novel WOA-based convolutional neural network classifier. *J. Integr. Sci. Technol.* **2023**, 11 (3), 523.