MULTIMODAL BRAIN TUMOR SEGMENTATION AND BEYOND

EDITED BY: Bjoern Menze and Spyridon Bakas
PUBLISHED IN: Frontiers in Computational Neuroscience and Frontiers in Neuroscience

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-170-3
DOI 10.3389/978-2-88971-170-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MULTIMODAL BRAIN TUMOR SEGMENTATION AND BEYOND

Topic Editors:
Bjoern Menze, Technical University of Munich, Germany
Spyridon Bakas, University of Pennsylvania, United States

Citation: Menze, B., Bakas, S., eds. (2021). Multimodal Brain Tumor Segmentation and Beyond. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-170-3

Table of Contents

frontiers in Computational Neuroscience

BRIEF RESEARCH REPORT
published: 12 July 2019
doi: 10.3389/fncom.2019.00044

Inception Modules Enhance Brain Tumor Segmentation

Daniel E. Cahall[1], *Ghulam Rasool*[1]*, *Nidhal C. Bouaynaya*[1] and *Hassan M. Fathallah-Shaykh*[2]

[1] Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ, United States, [2] Departments of Neurology and Mathematics, University of Alabama at Birmingham, Birmingham, AL, United States

Magnetic resonance images of brain tumors are routinely used in neuro-oncology clinics for diagnosis, treatment planning, and post-treatment tumor surveillance. Currently, physicians spend considerable time manually delineating different structures of the brain. Spatial and structural variations, as well as intensity inhomogeneity across images, make the problem of computer-assisted segmentation very challenging. We propose a new image segmentation framework for tumor delineation that benefits from two state-of-the-art machine learning architectures in computer vision, i.e., Inception modules and U-Net image segmentation architecture. Furthermore, our framework includes two learning regimes, i.e., learning to segment intra-tumoral structures (necrotic and non-enhancing tumor core, peritumoral edema, and enhancing tumor) or learning to segment glioma sub-regions (whole tumor, tumor core, and enhancing tumor). These learning regimes are incorporated into a newly proposed loss function which is based on the Dice similarity coefficient (DSC). In our experiments, we quantified the impact of introducing the Inception modules in the U-Net architecture, as well as, changing the objective function for the learning algorithm from segmenting the intra-tumoral structures to glioma sub-regions. We found that incorporating Inception modules significantly improved the segmentation performance ($p < 0.001$) for all glioma sub-regions. Moreover, in architectures with Inception modules, the models trained with the learning objective of segmenting the intra-tumoral structures outperformed the models trained with the objective of segmenting the glioma sub-regions for the whole tumor ($p < 0.001$). The improved performance is linked to multiscale features extracted by newly introduced Inception module and the modified loss function based on the DSC.

Keywords: gliomas, brain tumor segmentation, fully convolutional neural network, inception, U-net

OPEN ACCESS

Edited by:
Spyridon Bakas, University of Pennsylvania, United States

Reviewed by:
Jan Egger, Graz University of Technology, Austria
Siddhesh Pravin Thakur, University of Pennsylvania, United States

***Correspondence:**
Ghulam Rasool
rasool@rowan.edu

Received: *30 April 2019*
Accepted: *24 June 2019*
Published: *12 July 2019*

Citation:
Cahall DE, Rasool G, Bouaynaya NC and Fathallah-Shaykh HM (2019) Inception Modules Enhance Brain Tumor Segmentation. Front. Comput. Neurosci. 13:44. doi: 10.3389/fncom.2019.00044

1. INTRODUCTION

In recent years, there has been a proliferation of machine and especially deep learning techniques in the medical imaging field (Litjens et al., 2017). Deep learning algorithms also referred to as deep neural networks, are built using large stacks of individual artificial neurons, each of which performs primitive mathematical operations of multiplication, summation, and thresholding. One of the key reasons for the success of these modern deep neural networks is the idea of representation learning; the process of learning useful features automatically from the data as opposite to manual selection by expert humans (LeCun et al., 2015). Specifically, a convolutional neural network (CNN) is designed to extract features from two-dimensional grid data, e.g., images, through a series of

learned filters and non-linear activation functions. The set of features learned through this process can then be used to perform various downstream tasks such as image classification, object detection, and semantic or instance segmentation (LeCun et al., 2015).

Recently, U-Net (Ronneberger et al., 2015) which is an end-to-end fully convolutional network (FCN) (Long et al., 2015) was proposed for semantic segmentation of various structures in medical images. U-Net architecture is built using a contracting path, which captures high-resolution, contextual features while downsampling at each layer, and an expanding path, which increases the resolution of the output through upsampling at each layer (Ronneberger et al., 2015). The features from the contracting path are combined with features from the expanding path through skip connections (Drozdzal et al., 2016), ensuring localization of the extracted contextual features. Originally the U-Net was developed and applied to cell tracking, more recently the model has been applied to other medical segmentation tasks, such as, brain vessel segmentation (Livne et al., 2019), brain tumor segmentation (Dong et al., 2017), and retinal segmentation (Girard et al., 2019). Architectural variations and extensions of the U-Net algorithm, such as 3D U-Net (Kamnitsas et al., 2017; Sandur et al., 2018), H-DenseUNet (Li et al., 2018), RIC-UNet (Zeng et al., 2019), and Bayesian U-Net (Orlando et al., 2019) have been developed to tackle different segmentation problems in the medical imaging community.

Accurate semantic segmentation depends on the extraction of local structural as well as global contextual information from medical images during the learning process (training). Therefore, various multi-path architectures have been proposed in the medical image segmentation literature which extract information from given data at multiple scales (Havaei et al., 2017; Kamnitsas et al., 2017; Salehi et al., 2017). The concept of extracting and aggregating features at various scales has also been accomplished by Inception modules (Szegedy et al., 2015). However, the mechanism of feature extraction is different compared to multi-path architectures (Havaei et al., 2017; Kamnitsas et al., 2017; Salehi et al., 2017). Each Inception module applies filters of various sizes at each layer and concatenates resulting feature maps (Szegedy et al., 2015). The dilated residual Inception (DRI) block introduced in Shankaranarayana et al. (2019) was designed to accomplish multi-scale feature extraction in an end-to-end, fully convolutional retinal depth estimation model. The MultiResUNet recently proposed in Ibtehaz and Rahman (2019) combined a U-Net with residual Inception modules for multi-scale feature extraction; authors applied their architecture to several multimodal medical imaging datasets. Integrating Inception modules in a U-Net architecture has also been evaluated in the context of left atrial segmentation (Wang et al., 2019). An architecture proposed in Li and Tso (2018) for liver and tumor segmentation also incorporated inception modules, along with dilated Inception modules, in a U-Net. Concurrently and independently of this work, inception modules within U-Net have also been recently proposed for brain tumor segmentation in Li et al. (2019). However, authors used a cascade approach, i.e., first learn the whole tumor, then learn the tumor core, and finally learn the enhancing tumor, which requires three different models. Our proposed architecture is an end-to-end implementation with respect to all tumor subtypes.

The Multimodal Brain Tumor Image Segmentation (BRATS) challenge, started in 2012, has enabled practitioners and machine learning experts to develop and evaluate approaches on a continuously growing multi-class brain tumor segmentation benchmark (Menze et al., 2014). Based on the annotation protocol, deep learning architectures designed for the problem typically derive the segmentation using a pixel-wise softmax function on the output feature map (Isensee et al., 2018a). The softmax function enforces mutual exclusivity, i.e., a pixel can only belong to one of the intra-tumoral structures. The individual output segments are then combined to create the glioma sub-regions. Learning the glioma sub-regions directly using a pixel-wise sigmoid function on the output feature map has been discussed in Isensee et al. (2018b), as well as in Wang et al. (2018) using a cascaded approach.

In this work, we introduce an end-to-end brain tumor segmentation framework which utilizes a modified U-Net architecture with Inception modules to accomplish multi-scale feature extraction. Moreover, we evaluate the impact of training various models to segment the glioma sub-regions directly rather than the intra-tumoral structures. Both learning regimes were incorporated into a new loss function based on the Dice similarity Coefficient (DSC).

2. METHODS

2.1. Data and Preprocessing

All experiments were conducted on the BRATS 2018 dataset (Menze et al., 2014; Bakas et al., 2017a,b,c, 2018), which consists of magnetic resonance imaging (MRI) data of 210 high-grade glioma (HGG) and 75 low-grade glioma (LGG) patients. Each patient's MRI data contained four MRI sequences: T2-weighted (T2), T1, T1 with gadolinium enhancing contrast (T1C), and Fluid-Attenuated Inversion Recovery (FLAIR) images. Furthermore, pixel-level manual segmentation markings are provided in the BRATS dataset for three *intra-tumoral* structures: necrotic and non-enhancing tumor core (label = 1), peritumoral edema (label = 2), and enhancing tumor (label = 4). For the intra-tumoral structures, following *glioma sub-regions* (Menze et al., 2014) were defined: whole tumor (WT) which encompasses all three intra-tumoral structures (i.e., label = $1 \cup 2 \cup 4$), tumor core (TC) that contains all but the peritumoral edema (i.e., label = $1 \cup 4$), and enhancing tumor (ET) (label = 4). Different sequences provide complementary information for identifying the intra-tumoral structures: FLAIR highlights the peritumoral edema, T1C distinguishes the ET, and T2 highlights the necrotic and non-enhancing tumor core. Converting from the intra-tumoral structures to the glioma sub-regions is a linear, reversible transformation; the glioma sub-regions are generated from the intra-tumoral structures, and provided the glioma sub-regions, the original intra-tumoral structures can be recovered.

The BRATS dataset is provided in a preprocessed format, i.e., all the images are skull-stripped, resampled to an isotropic 1 mm^3 resolution, and all four modalities of each patient are co-registered. We performed additional preprocessing that

included (in order): (1) obtaining the bounding box of the brain in each image, and extracting the selected portion of the image, effectively zooming in on the brain and disregarding excess background pixels, (2) re-sizing the cropped image to 128 x 128 pixels, (3) removing images which contained no tumor regions in the ground truth segmentation, (4) applying an intensity windowing function to each image such that the lowest 1% and highest 99% pixels were mapped to 0 and 255, respectively, and (5) normalizing all images by subtracting the mean and dividing by the standard deviation of the dataset.

2.2. Segmentation Model Architecture

We propose a new architecture based on the 2D U-Net and factorized convolution Inception module (Ronneberger et al., 2015; Szegedy et al., 2016). Each convolutional layer in the original U-Net was replaced with an Inception module that included multiple sets of 3 × 3 convolutions, 1 × 1 convolutions, 3 × 3 max pooling, and cascaded 3 × 3 convolutions. A cartoon of the proposed network architecture with an expanded view of the Inception module is presented in **Figure 1**. We note that at each layer on the contracting path, the height and width of the feature maps are halved and the depth is doubled until reaching the bottleneck i.e., the center of the "U." Conversely, on the expanding path, the height and width of the feature maps are doubled and the depth is halved at each layer until reaching the output (i.e., segmentation mask for the given input image). Furthermore, each set of feature maps generated on the contracting path are concatenated to the corresponding feature maps on the expanding path. We used rectified linear

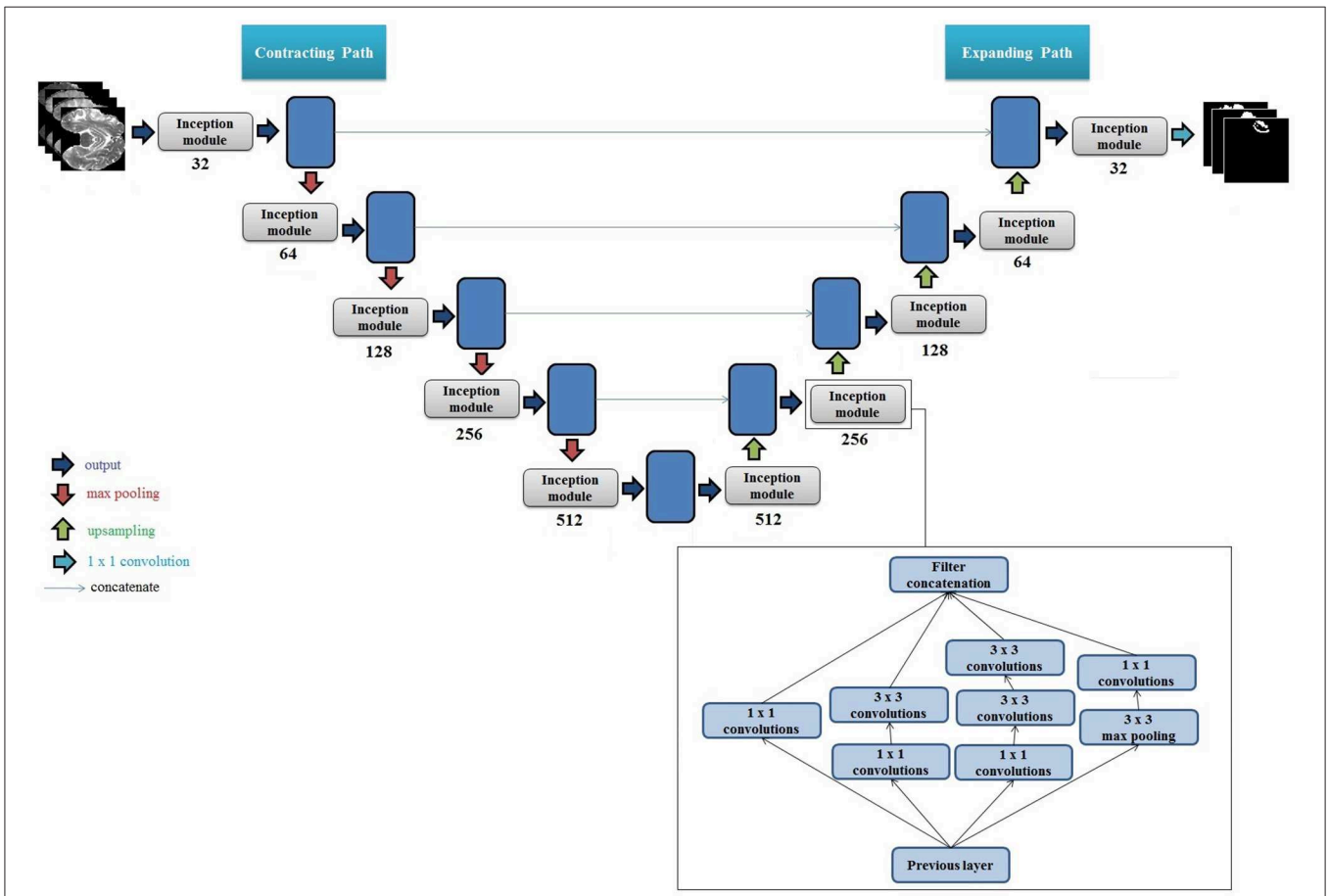


FIGURE 1 | Cartoon of the proposed segmentation architecture. The set of numbers shown below each Inception module indicate total number of filters used, and height, width, and depth of the input feature map. The number of filters at each layer double on the encoder side, and the size of the output feature map (height and width) halve. The multiplication by 4 for each depth value is due to the 4 filter variations in the Inception module, which generates 4 sets of equally sized feature maps that are concatenated. The feature maps are then downsampled using max pooling, which halves their height and width. This process is repeated until reaching the bottleneck i.e., the "center" of the U. Upsampling is then performed which doubles the height and width of each feature map, and the feature maps from the corresponding stage on the contracting path are concatenated to the upsampled feature maps (shown by blue lines). The concatenation of the feature maps from the contracting path doubles the depth of the output feature map on the expanding path, hence the multiplication by 8. At the last layer on the expanding path, the output height and width are equivalent to the height and width of the original input images. A set of 1 × 1 convolutions is then applied to reduce the depth of the last feature map to equal the number of classes (tumor regions). A pixel-wise activation function is then applied to then convert the reduced feature map to binary segmentation images. Right Bottom: Internal architecture of one Inception module with multiple convolutional filters and max pooling filters is presented. The numbers in each block represent convolution filter size. We used two 3 × 3 filters in series to get an equivalent receptive field of a 5 × 5 convolutional filter.

unit (ReLU) as the activation function for each layer, and performed batch normalization (Ioffe and Szegedy, 2015) in each Inception module.

The input to our model is an $N \times M \times D$ pixel image and the output of the model is an $N \times M \times K$ tensor. In out settings, $N = M = 128$ pixels, $D = 4$ which represents all four MRI modalities, and $K = 3$ which represents total number of segmentation classes, i.e., intra-tumoral structures or the glioma sub-regions. Each slice of K is a binary image representing the predicted segments for the ith class where $0 \leq i \leq K-1$. The binary images are generated by pixel-wise activation functions, i.e., sigmoid for glioma sub-regions and softmax for intra-tumoral structures.

2.3. Evaluation Metric and Objective (Loss) Function

Dice Similarity Coefficient (DSC) is extensively used for the evaluation of segmentation algorithms in medical imaging applications (Bakas et al., 2017a). The DSC between a predicted binary image P and a ground truth binary image G, both of size $N \times M$ is given by:

$$DSC(P, G) = 2\frac{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P_{ij}G_{ij}}{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P_{ij} + \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} G_{ij}}, \quad (1)$$

where i and j represent pixel indices for the height N and width M. The range of DSC is $[0, 1]$, and a higher value of DSC corresponds to a better match between the predicted image P and the ground truth image G.

Our objective function (or the loss function) for the proposed learning algorithm consisted of a modified version of DSC (Equation 1). Specifically, following modification were made: (1) we changed the sign of the DSC coefficient to formulate a standard deep learning optimization (minimization) problem, (2) introduced log function, and (3) introduced a new parameter γ to cater for extremely large values of the loss function. For example, if a ground truth segment had very few white pixels $\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} G_{ij} \approx 0$, the model may predict no white pixels $\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P_{ij} = 0$ resulting in an extremely large loss function. In our preliminary experiments, we found empirically that $\gamma = 100$ provided the best segmentation performance. The resulting expression for the loss function is given as:

$$\mathcal{L}_{DSC}(P, G) = -\log\left[2\frac{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P_{ij}G_{ij} + \gamma}{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P_{ij} + \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} G_{ij} + \gamma}\right]. \quad (2)$$

The loss function presented in Equation (2) is able to handle binary cases only (e.g., tumor and not tumor). The same can be extended for the multi-class cases as:

$$\mathcal{L}_{DSC}(P, G) = -\log\left[\frac{1}{K}\sum_{i=0}^{K-1} DSC(P_i, G_i)\right], \quad (3)$$

where K is the total number of classes.

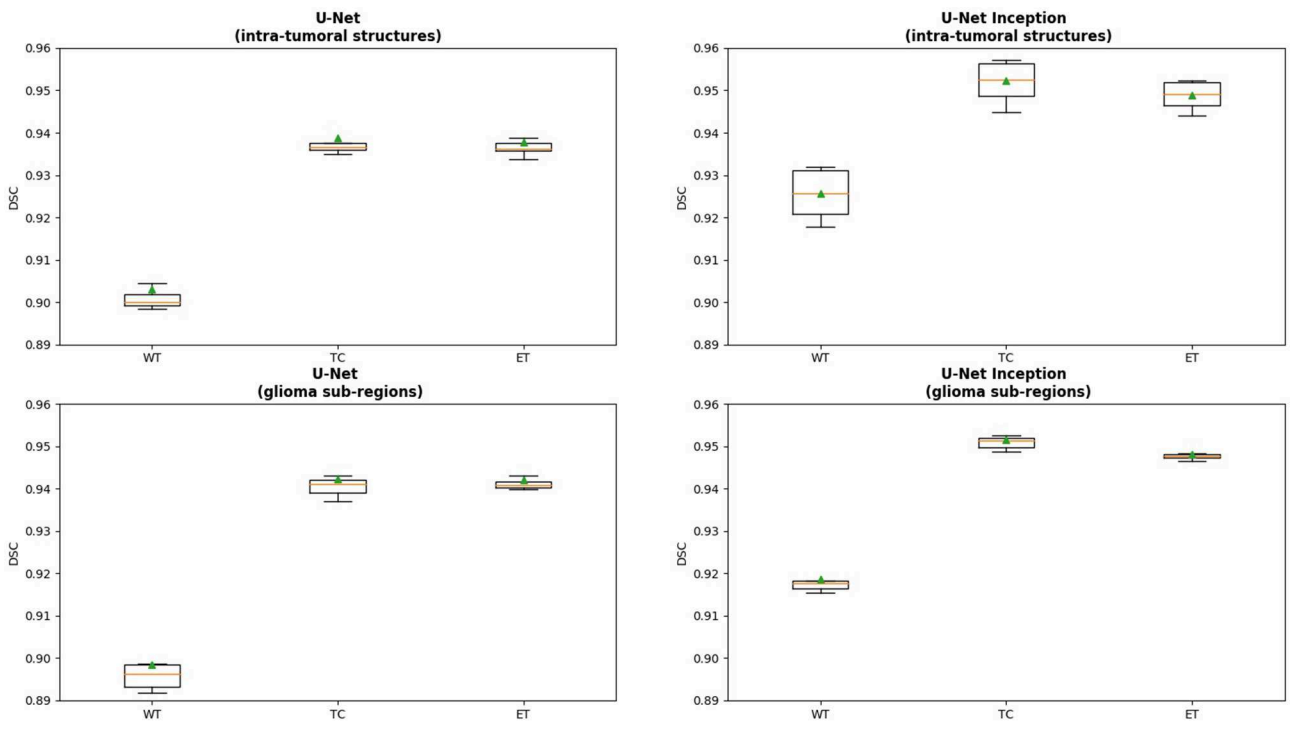


FIGURE 2 | Box plot displaying the results for each model variation. The x-axis is the glioma sub-region, and the y-axis is the DSC. The median value is denoted by the horizontal orange line, and the mean is denoted by the green triangle. Abbreviations used are: WT, Whole Tumor; TC, Tumor Core; and ET, Enhancing Tumor.

2.4. Experimental Setup and Model Training

We performed an ablation study to quantify the effects of introducing Inception modules in the U-Net architecture as well as the impact of different segmentation objectives, i.e., learning to segment intra-tumoral structures or glioma sub-regions. Specifically, we trained four different models, i.e., two variations of the U-Net architecture (with intra-tumoral structures and glioma sub-regions) and two variations of the U-Net with Inception module (intra-tumoral structures and glioma sub-regions).

We trained all four models under same conditions to ensure consistency and a fair comparison. All four models were trained using k-fold cross-validation. The dataset was randomly split into k mutually exclusive subsets of equal or near equal size. Each algorithm was run k times subsequently, each time taking one of the k splits as the validation set and the rest as the training

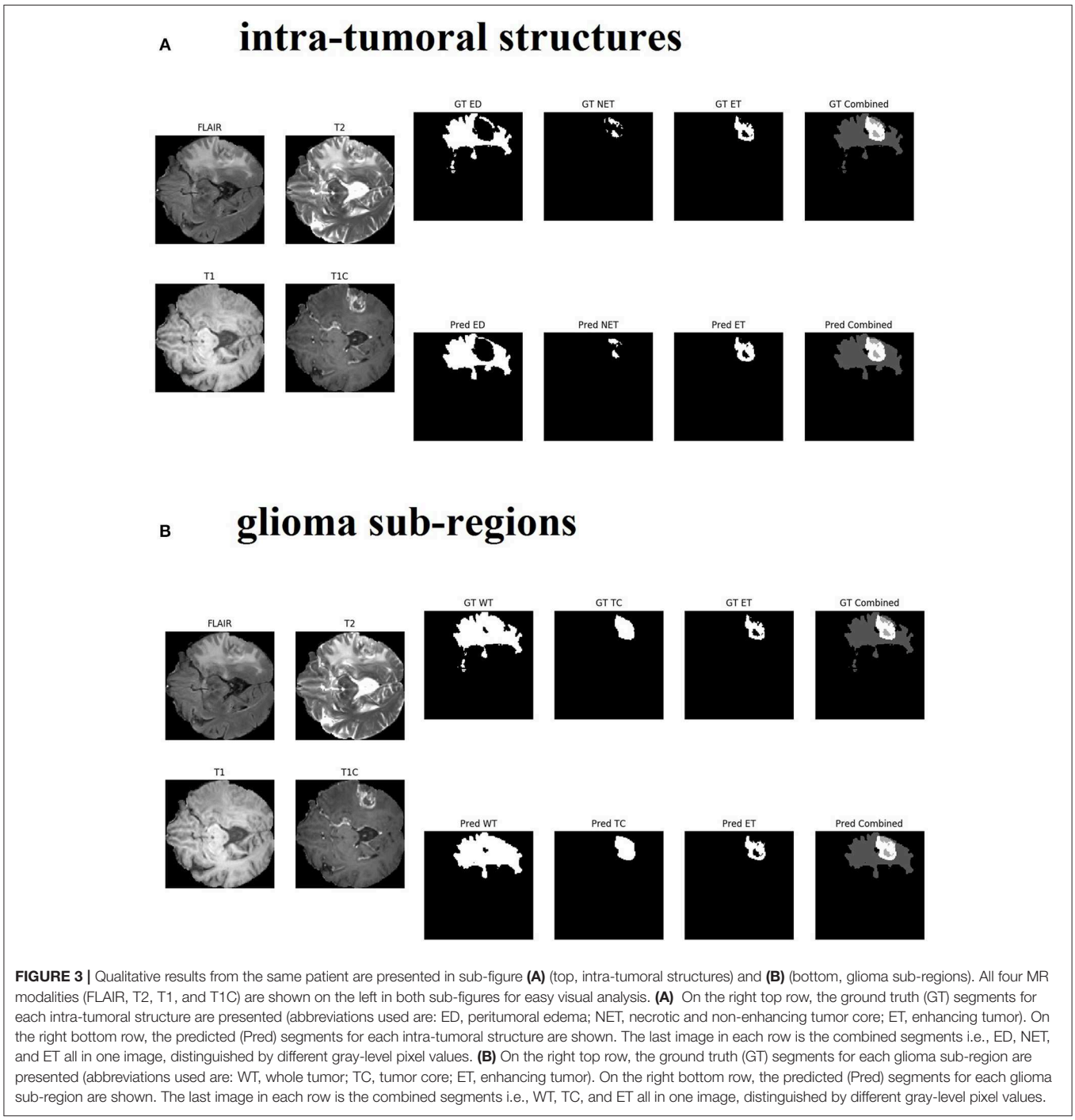


FIGURE 3 | Qualitative results from the same patient are presented in sub-figure **(A)** (top, intra-tumoral structures) and **(B)** (bottom, glioma sub-regions). All four MR modalities (FLAIR, T2, T1, and T1C) are shown on the left in both sub-figures for easy visual analysis. **(A)** On the right top row, the ground truth (GT) segments for each intra-tumoral structure are presented (abbreviations used are: ED, peritumoral edema; NET, necrotic and non-enhancing tumor core; ET, enhancing tumor). On the right bottom row, the predicted (Pred) segments for each intra-tumoral structure are shown. The last image in each row is the combined segments i.e., ED, NET, and ET all in one image, distinguished by different gray-level pixel values. **(B)** On the right top row, the ground truth (GT) segments for each glioma sub-region are presented (abbreviations used are: WT, whole tumor; TC, tumor core; ET, enhancing tumor). On the right bottom row, the predicted (Pred) segments for each glioma sub-region are shown. The last image in each row is the combined segments i.e., WT, TC, and ET all in one image, distinguished by different gray-level pixel values.

set. In our experiments, we set $k = 10$, which means that each model was trained 10 times using a different set of 90% of the data and validated on the remaining 10% data. In total, our experimental setup generated 40 models, i.e., 10 variations per model. Later, mean and standard deviation (SD) were calculated and are reported for each model in the Results section.

We used stochastic gradient descent with an adaptive moment estimator (Adam) for training all models and their variations (Kingma and Ba, 2014). The initial learning rate was set to 10^{-4} which was exponentially decayed every 10 epochs. The batch size was set to 64 and each model was trained for 100 epochs. All learnable parameters, i.e., weights and biases of the models were initialized based on the He initialization method (He et al., 2015). The Keras (Chollet et al., 2015) application programming interface (API) with TensorFlow (Abadi et al., 2016) backend was used for implementation of all models. All models were trained on a Google Cloud Compute instance with 4 NVIDIA TESLA P100 graphical processing units (GPUs).

2.5. Model Testing and Statistical Analysis of Results

After training, each model was tested on the entire BRATS 2018 dataset. For the models which learned to segment the intra-tumoral structures, the predicted intra-tumoral structure segments were combined to produce the glioma sub-regions, and DSC for each glioma sub-region was computed. For models which learned to segment the glioma sub-regions directly, DSC values were readily computed. The process was repeated for each image, and after evaluating all images, the average DSC score was calculated for each glioma sub-region. Overall, the process resulted in 4 sets of 10 DSC scores, one for each glioma sub-region. All four models were compared for statistical significance using a two-tailed Student's t-test with equal variance and with the probability of Type-I error set to $\alpha = 0.05$.

3. RESULTS

We present cross-validation DSC for all four models that were trained and tested on the BRATS 2018 dataset. In **Figure 2**, we provide a box plot for each model variation. The glioma sub-region is on the x-axis and the DSC is on the y-axis for each plot. We note that for intra-tumoral structures, adding Inception modules to the U-Net resulted in statistically significant improvements in WT (DSC improved from 0.903 to 0.925, $p < 0.001$), TC (0.938 to 0.952, $p < 0.001$), and ET (0.937 to 0.948, $p < 0.001$). Similarly, for the glioma sub-regions, adding Inception modules to the U-Net also resulted in statistically significant improvements in WT (0.898 to 0.918, $p < 0.001$), TC (0.942 to 0.951, $p = 0.001$), and ET (0.942 to 0.948, $p = 0.002$).

Changing the objective from learning the intra-tumoral structures to learning the glioma sub-regions in the U-Net resulted in no difference in performance for WT (0.903 to 0.898, $p = 0.307$), TC (0.938 to 0.942, $p = 0.284$), and ET (0.937 to 0.942, $p = 0.098$). However, U-Net with Inception modules which learned the intra-tumoral structures outperformed U-Net with Inception modules which learned the glioma sub-regions in WT (0.918 to 0.925, $p = 0.007$), but there was no performance difference for TC (0.952 to 0.951, $p = 0.597$) and ET (0.948 to 0.948, $p = 0.402$). Qualitative results on the same patient from a U-Net with Inception modules which learned the intra-tumoral structures and U-Net with Inception modules which learned the glioma sub-regions are presented in **Figures 3A,B**, respectively. In **Table 1**, we provide a summary of statistical comparisons, i.e., p-values from Student's t-test performed to compare different models. Statistically significant p-values are in shown bold font.

TABLE 1 | Results of statistical comparison, i.e., p-values from two-tailed t-tests comparing the models in the first column with the models in the second columns.

Model 1	Model 2	*p*-values		
		WT	**TC**	**ET**
U-Net intra-tumoral structures	U-Net glioma sub-regions	0.307	0.284	0.098
	U-Net Inception intra-tumoral structures	**<0.001**	**<0.001**	**<0.001**
U-Net Inception glioma sub-regions	U-Net glioma sub-regions	**<0.001**	**0.001**	**0.002**
	U-Net Inception intra-tumoral structures	**0.007**	0.597	0.402

Statistically significant p-values are present in bold font.

4. DISCUSSION AND CONCLUSIONS

We set out to tackle the challenging problem of pixel-level segmentation of brain tumors using MRI data and deep learning models. We introduced a new framework building on well-known U-Net architecture and Inception modules. We explored two different learning objectives: (1) learning to segment glioma sub-regions (WT, TC, and ET), and (2) learning to segment intra-tumoral structures (necrotic and non-enhancing tumor core, peritumoral edema, and enhancing tumor). Both learning objectives were incorporated into the newly proposed DSC based loss function. Our framework resulted into four different model variations, i.e., (1) a U-Net with learning objective of intra-tumoral structures, (2) U-Net with glioma sub-regions, (3) U-Net with Inception module and intra-tumoral structures, and finally (4) U-Net with Inception module and learning objective of glioma sub-regions.

We found that integrating Inception modules in the U-Net architecture resulted in statistically significant improvement in tumor segmentation performance that was quantified using k-fold cross-validation ($p < 0.05$ for all three glioma sub-regions). We consider that the observed improvement in the validation accuracy is linked to multiple convolutional filters of different sizes employed in each Inception module. These filters are able to capture and retain contextual information at multiple scales during the learning process, both in the contracting as well as expanding paths. We also consider that the improvement in the tumor segmentation accuracy is linked to the new loss function based on the modified DSC (i.e., Equation 3). In our proposed framework, we evaluate our models using DSC and the learning objective or the loss function (Equation 3) used

for training these algorithms is also based on DSC. This is in contrast with conventional deep learning paradigms being used in natural image segmentation, such as, Mask R-CNN, where the loss function is based on multi-class cross-entropy and the evaluation metric is based on Intersection-over-Union (IoU) or DSC score (He et al., 2017). Furthermore, our DSC scores for each glioma sub-region on the BRATS 2018 training dataset are comparable or exceed the results of other recent published architectures such as the No New-Net, which achieved second place in the BRATS 2018 competition (Isensee et al., 2018b), and the ensemble approach proposed in Kao et al. (2018).

Our results also demonstrate that changing the learning objective from intra-tumoral structures to glioma sub-regions in the architectures with Inception modules produced a statistically significant positive impact only on WT, while not affecting TC and ET. Since the only difference between TC and WT is the peritumoral edema, these results suggest that learning to segment the peritumoral edema independently is more effective than learning in context of other two intra-tumoral structures. We hypothesize that learning to segment WT directly may be difficult for the model because it requires extracting information from multiple modalities (T1, T1C, T2, and FLAIR); however, the segmentation of peritumoral edema alone can primarily be learned from FLAIR data. Therefore, for the proposed framework, we recommend using intra-tumoral structures for learning with U-Net Inception architecture.

DATA AVAILABILITY

The BRATS 2018 training dataset analyzed for this study can be found in the Image Processing Portal of the CBICA@UPenn [https://ipp.cbica.upenn.edu/].

AUTHOR CONTRIBUTIONS

The architecture was conceived by DC. The experiments were designed by GR, NB, and HF-S. The data was analyzed by HF-S and DC conducted the experiments and wrote the manuscript with support from GR, NB, and HF-S. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

FUNDING

This work was supported by the U.S. Department of Education Graduate Assistance in Areas of National Need (GAANN) Grant Number P200A180055 and the U.S. National Science Foundation (NSF) Award DUE-1610911.

ACKNOWLEDGMENTS

The authors would like to acknowledge Google Cloud Platform (GCP) for their computational resources.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "TensorFlow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA), 265–283.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). *Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection.* The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). *Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection.* The Cancer Imaging Archive.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv [Preprint]*.

Chollet, F. et al. (2015). *Keras*. Available online at: https://keras.io

Dong, H., Yang, G., Liu, F., Mo, Y., and Guo, Y. (2017). "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Communications in Computer and Information Science Medical Image Understanding and Analysis* (Edinburgh, UK), 506–517.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications Lecture Notes in Computer Science* (Athens), 179–187.

Girard, F., Kavalec, C., and Cheriet, F. (2019). Joint segmentation and classification of retinal arteries/veins from fundus images. *Artif. Intell. Med.* 94, 96–109. doi: 10.1016/j.artmed.2019.02.004

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving Deep into Rectifiers: surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago).

Ibtehaz, N., and Rahman, M. S. (2019). MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. *arXiv [Preprint]*.

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv [Preprint]*.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018a). "Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries Lecture Notes in Computer Science* (Quebec City, QC), 287–297.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018b). "No New-Net," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 234–244.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kao, P.-Y., Ngo, T., Zhang, A., Chen, J. W., and Manjunath, B. (2018). "Brain tumor segmentation and tractographic feature extraction from structural MR images for overall survival prediction," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 128–141.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv [Preprint]*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436. doi: 10.1038/nature14539

Li, H., Li, A., and Wang, M. (2019). A novel end-to-end brain tumor segmentation method using improved fully convolutional networks. *Comput. Biol. Med.* 108, 150–160. doi: 10.1016/j.compbiomed.2019.03.014

Li, S., and Tso, G. K. F. (2018). Bottleneck Supervised U-Net for Pixel-wise Liver and Tumor Segmentation. *arXiv [Preprint]*.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* 37, 2663–2674. doi: 10.1109/TMI.2018.2845918

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Livne, M., Rieger, J., Aydin, O. U., Taha, A. A., Akay, E. M., Kossen, T., et al. (2019). A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. *Front. Neurosci.* 13:97. doi: 10.3389/fnins.2019.00097

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA).

Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Orlando, J. I., Seeböck, P., Bogunović, H., Klimscha, S., Grechenig, C., Waldstein, S., et al. (2019). U2-Net: a Bayesian U-Net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological OCT scans. *arXiv [Preprint]*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (Munich), 234–241.

Salehi, S. S. M., Erdogmus, D., and Gholipour, A. (2017). Auto-context convolutional neural network (Auto-Net) for brain extraction in magnetic resonance imaging. *IEEE Trans. Med. Imaging* 36, 2319–2330. doi: 10.1109/TMI.2017.2721362

Sandur, P., Naveena, C., Aradhya, V. M., and Nagasundara, K. B. (2018). Segmentation of brain tumor tissues in HGG and LGG MR images using 3D U-Net convolutional neural network. *Int. J. Nat. Comput. Res.* 7, 18–30. doi: 10.4018/IJNCR.2018040102

Shankaranarayana, S. M., Ram, K., Mitra, K., and Sivaprakasam, M. (2019). Fully convolutional networks for monocular retinal depth estimation and optic disc-cup segmentation. *IEEE J. Biomed. Health Inform.* Available online at: https://ieeexplore.ieee.org/document/8642288

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV).

Wang, C., Rajchl, M., Chan, A., and Ukwatta, E. (2019). "An ensemble of U-Net architecture variants for left atrial segmentation," in *Medical Imaging 2019: Computer-Aided Diagnosis*, Vol. 10950 (International Society for Optics and Photonics) (San Diego, CA), 109500M.

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2018). "Automatic Brain Tumor Segmentation Using Cascaded Anisotropic Convolutional Neural Networks," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries Lecture Notes in Computer Science* (Quebec City, QC), 178–190.

Zeng, Z., Xie, W., Zhang, Y., and Lu, Y. (2019). RIC-Unet: an improved neural network based on Unet for nuclei segmentation in histology images. *IEEE Access* 7, 21420–21428. doi: 10.1109/ACCESS.2019.2896920

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 30 July 2019
doi: 10.3389/fncom.2019.00052

Prediction of 1p/19q Codeletion in Diffuse Glioma Patients Using Pre-operative Multiparametric Magnetic Resonance Imaging

*Donnie Kim[1], Nicholas Wang[2], Viswesh Ravikumar[1], D. R. Raghuram[1], Jinju Li[2], Ankit Patel[1], Richard E. Wendt III[1], Ganesh Rao[1] and Arvind Rao[2]**

[1] Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX, United States, [2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States

This study compared the predictive power and robustness of texture, topological, and convolutional neural network (CNN) based image features for measuring tumors in MRI. These features were used to predict 1p/19q codeletion in the MICCAI BRATS 2017 challenge dataset. Topological data analysis (TDA) based on persistent homology had predictive performance as good as or better than texture-based features and was also less susceptible to image-based perturbations. Features from a pre-trained convolutional neural network had similar predictive performances and robustness as TDA, but also performed better using an alternative classification algorithm, k-top scoring pairs. Feature robustness can be used as a filtering technique without greatly impacting model performance and can also be used to evaluate model stability.

Keywords: multiparametric MRI, image perturbation, radiomic features, glioma, persistent homology, 1p/19q codeletion

OPEN ACCESS

Edited by:
Spyridon Bakas, University of Pennsylvania, United States

Reviewed by:
Ken Chang, Massachusetts Institute of Technology, United States
Carlos Alberto Silva, University of Minho, Portugal

***Correspondence:**
Arvind Rao
ukarvind@umich.edu

Received: *30 April 2019*
Accepted: *11 July 2019*
Published: *30 July 2019*

Citation:
Kim D, Wang N, Ravikumar V, Raghuram DR, Li J, Patel A, Wendt RE III, Rao G and Rao A (2019) Prediction of 1p/19q Codeletion in Diffuse Glioma Patients Using Pre-operative Multiparametric Magnetic Resonance Imaging. Front. Comput. Neurosci. 13:52. doi: 10.3389/fncom.2019.00052

BACKGROUND

1p/19q codeletion, is a genetic loss event that is somewhat rare in gliomas (Fuller and Perry, 2005; Eckel-Passow et al., 2015). It involves the complete deletion of the short arm of chromosome 1 alongside the deletion of the long arm of chromosome 19. Patients with this genetic loss event have been shown to have markedly improved prognosis and overall survival as compared to patients without 1p/19q codeletion (Boots-Sprenger et al., 2013; Cairncross et al., 2013; Van M den et al., 2013). The ability to identify patients from radiologic imaging would help to tailor treatment for this subtype of brain cancer.

Radiomics is the study of tumor imaging data, and the use of the imaging features to predict prognosis or genetic markers of these tumors. Radiological studies are standard of care for most cancer patients, but genetic profiling is available only for a subset of cancer patients (Gillies et al., 2015). Thus, understanding the relationship between tumor appearance on magnetic resonance imaging (MRI) and the genetic profile of a tumor could help to predict prognosis or to subtype tumors and thereby deliver more precise care to larger patient populations.

A number of publicly available datasets and toolkits exist for measuring texture-based features on tumors (Clark et al., 2013; van Griethuysen et al., 2017). However, while there has been progress in measuring these features, there is some concern about the robustness and generalizability of radiomic features. Other studies on CT scans have shown that some texture-based features are

not stable under perturbation in test-retest comparisons (Bogowicz et al., 2016; van Timmeren et al., 2016). In order further to assess the degree of instability, this study has investigated the effect of image perturbations on additional feature types beyond texture, and their eventual effect on classification power in MRI scans.

METHODS

A set of brain MRI data were drawn from the MICCAI BRATS 2017 challenge dataset (Menze et al., 2015; Bakas et al., 2017a, 2018). The multimodal Brain Tumor Image Segmentation Benchmark (BRATS) 2017 dataset was originally designed for the brain tumor segmentation challenge and comprises pathologically confirmed LGG ($n = 65$) and HGG ($n = 102$) cases from The Cancer Imaging Archive (TCIA) (Bakas et al., 2017b,c). The dataset contains pre-operative multimodal MRI sequences, namely T1, T1-post, T2, and FLAIR, and was acquired with differing imaging/clinical protocols and scanners from 19 different institutions. All tumor volumes in the imaging dataset had been segmented manually by one to four different experienced neuroradiologists.

Genetic markers for this TCIA dataset were gathered from The Cancer Genomics Archive (TCGA). The patients were first retrospectively identified with histologically confirmed WHO grade II-IV gliomas ($n = 1{,}122$) and their corresponding 1p/19q chromosome codeletion statuses (after surgical biopsy). In addition, the patients' age, gender, Karnofsky Performance Score (KPS) were collected as clinical variables.

These four sequences were co-registered to the T1 post-sequence as it had the highest spatial resolution. They were then resampled to $1 \times 1 \times 1$ mm isotropically in an axial orientation by using a linear interpolation algorithm. Then, all images were skull-stripped to anonymize the patient information and remove extraneous regions of the scan (Bauer et al., 2012).

The scans were prepared by performing N4 bias correction, normalizing intensity values by interquartile range, and cropping and reshaping to the volume of interest. Normalization of the intensity was performed based on the interquartile range for a particular modality of the non-tumor brain volume. The slices were resampled to a 142×142 image size that was cropped to the tumor area of interest. This methodology is similar to that used by Chang et al. (2018a) in order to provide the type of input that the neural network anticipated.

The breakdown of the dataset for 1p/19q codeletion vs. non-codeleted cases was heavily skewed toward the non-codeleted cases, with 13 cases with codeletion and 130 without codeletion. As such, the codeleted cases were heavily oversampled in slice selection at a 20:3 ratio to achieve a closer balance of class ratio. The largest 20 image patch slices for each codeleted scan was taken. For the non-codeleted scans the 50, 75, and 100th percentile slices (based on size) were taken.

The dataset was split patient-by-patient into sets of 80% for training and 20% for testing. This preserved the class ratio in the training and testing sets, as the number of positive cases was so low. This process was repeated 10 times independently for a total of 10 independent splits. Each of these independent splits had the entire analytic process performed to assess the robustness of the results. The training set was used in 5-fold cross-validation for each of the models, where patients were kept together in the cross-validation folds.

The three types of features measured in these scans were texture-based features, persistent homology topological features, and features based on a pre-trained convolutional neural network (**Figure 1**). The texture features were extracted slice-by-slice using the Pyradiomics package (van Griethuysen et al., 2017). The types of features were based on the tumor region of interest on each of the modalities. The texture features that were extracted included: first-order intensity features, shape features, gray-level co-occurrence matrix features (GLCM), gray-level run length matrix features (GLRLM), gray level size zone matrix features (GLSZM), and neighboring gray-tone difference matrix features (NGTDM).

It is well-known that MRI studies suffer from a variety of noise sources, so the underlying integrity of the image data carries some uncertainty. A topological approach was evaluated to see if the features generated were less susceptible to this uncertainty than traditional texture-based approaches. These topological features were based on persistent homology and how the topology changes with shifts in the image intensity threshold. Barcodes describe when a connected component or tunnel was created and destroyed by this shifting threshold (**Figure 2**; Adcock et al., 2014). These barcodes were collected with the GUDHI python package (Maria, 2015).

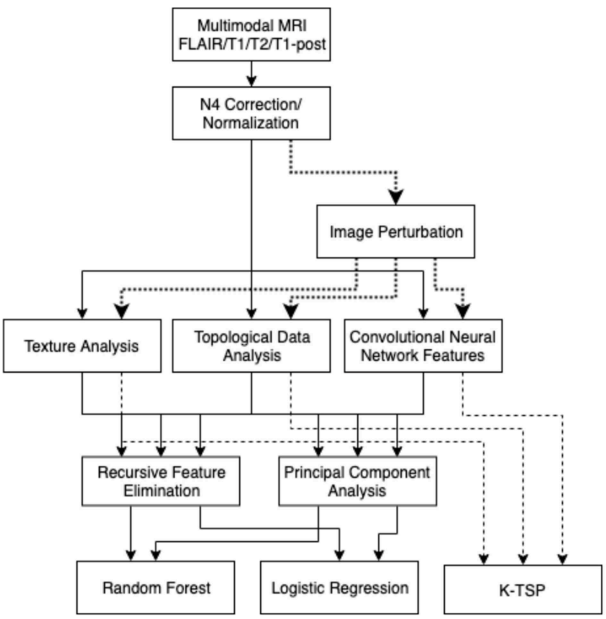


FIGURE 1 | Analysis pipeline: images are normalized, then the three types of features were collected. These features are filtered with RFE and PCA, then used to build a random forest model or logistic regression model. Image perturbations are used as an additional filter by including only relatively robust features. The kTSP algorithm used the same feature set to build its predictions.

These barcodes were characterized by their polynomial features, along with statistical features about their birth and death intensities, bar lengths, and death intensity distribution. These features were based on work in Adcock et al. (2013) and Giansiracusa et al. (2017).

A pre-trained convolutional neural network (CNN) was used to calculate deep learning-based features, from Chang's work on IDH1 mutation (Chang et al., 2018a). Chang's model was useful for this investigation as it focused on gliomas and featured the same MRI modalities as were present in this study (T1, T2, FLAIR, and T1-post). The second to last layer of the network was used to extract features rather than to feed into a softmax layer to predict IDH1 mutation. We expected the network to produce some features that are relevant to this 1p/19q dataset because the current work was of the same fundamental nature as the problem in Chang's work.

Two versions of feature reduction/selection were evaluated in the training set of this study: recursive feature elimination (RFE) and principal component analysis (PCA). RFE was performed with 10-fold cross-validation, to determine the optimal number of features (k), then the k best features were selected. PCA was performed and a cutoff of 95% cumulative variance was used to cull the insignificant components in PCA reduction.

Each of the feature sets—texture, topology, and CNN—had feature selection performed, and then those features were fed into a random forest model and a logistic regression model. The models were tuned using 5-fold cross-validation with folds that kept patients within the same fold. The random forest models were optimized over a number of hyperparameters including: tree counts of 200–2000, maximum depths of 10–100, the minimum sample split, and minimum leaf size. The logistic regression models had normalization hyperparameters of L1 vs. L2 normalization, and regularization strength from 10^{-3} to 10^{5}.

The models were evaluated primarily on the held-out 20% testing set, where area under the receiver operator curve (AUROC), accuracy, sensitivity, and specificity were measured. Additionally, combined models, which used features from

TABLE 1 | Test set statistics across 10 independent splits.

	AUROC	STD of AUROC	Sensitivity	Specificity	Accuracy
Texture only RF RFE	0.660	0.120	0.782	0.558	0.669
Texture only LR RFE	0.566	0.139	0.775	0.479	0.629
Texture only RF PCA	0.527	0.071	0.543	0.644	0.581
Texture only LR PCA	0.502	0.093	0.573	0.610	0.583
TDA only RF RFE	0.698	0.085	0.653	0.738	0.682
TDA only LR RFE	0.710	0.094	0.723	0.675	0.692
TDA only RF PCA	0.626	0.132	0.647	0.648	0.638
TDA only LR PCA	0.691	0.135	0.677	0.694	0.676
CNN only RF RFE	0.708	0.139	0.905	0.546	0.727
CNN only LR RFE	0.644	0.110	0.775	0.565	0.669
CNN only RF PCA	0.672	0.133	0.627	0.750	0.675
CNN only LR PCA	0.673	0.081	0.823	0.546	0.686
Combined RF RFE	0.689	0.150	0.877	0.552	0.714
Combined LR RFE	0.685	0.135	0.770	0.638	0.700
Combined RF PCA	0.612	0.148	0.655	0.637	0.638
Combined LR PCA	0.675	0.121	0.865	0.525	0.698
Clinical per patient RF	0.713	0.106	0.667	0.854	0.800
Clinical per patient LR	0.577	0.097	0.467	0.819	0.759

Darker blue indicates improved AUROC.

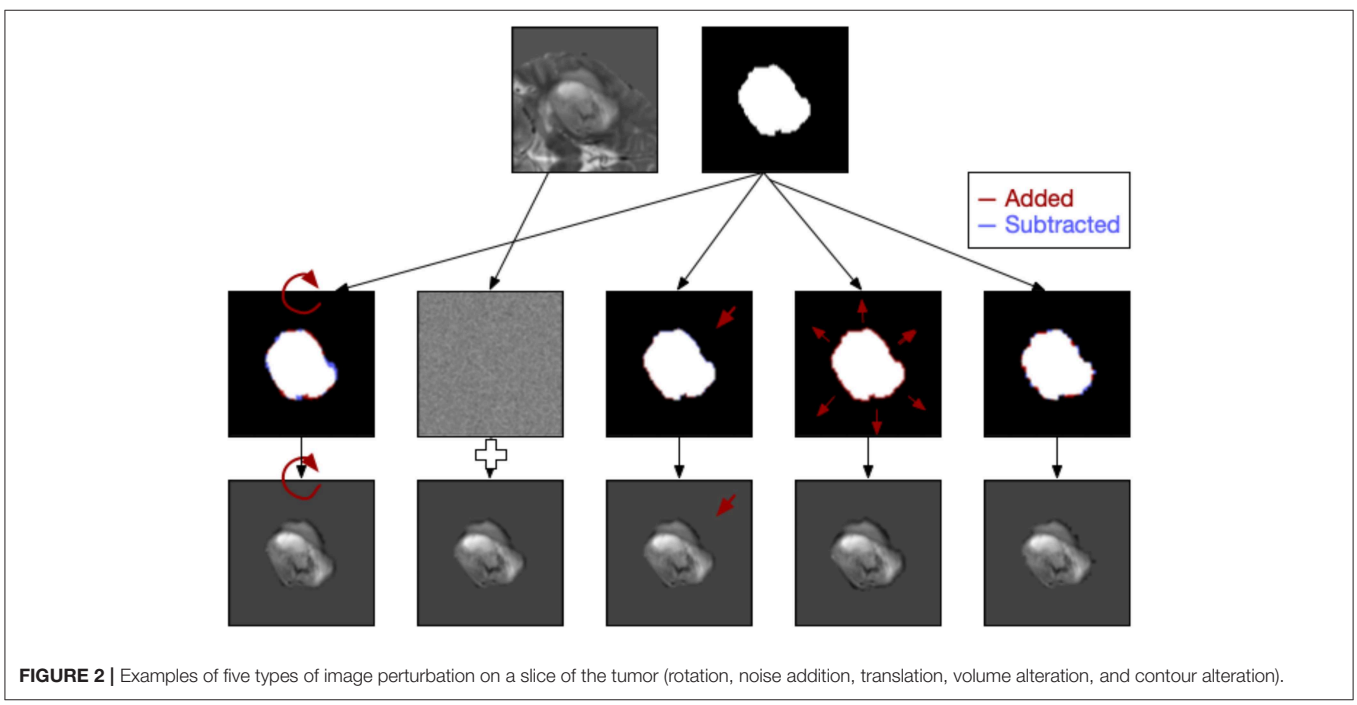


FIGURE 2 | Examples of five types of image perturbation on a slice of the tumor (rotation, noise addition, translation, volume alteration, and contour alteration).

texture, topology, and CNN, were also tested using the same approach. The clinical patient characteristics (age, sex, and Karnofsky performance score) were tested independently to gauge their performance in comparison to the imaging-based features.

Robustness of Features

Each of the image slices was perturbed using image processing techniques to produce relatively small changes to the image following the approach of Zwanenburg et al. (2019). Five classes of perturbation were performed on the images: image rotation (R), image translation (T), image Gaussian white noise addition (N), mask volume alteration (V), and contour randomization (C). Images and masks were rotated around the mask center of mass to approximate changes in head position in the scanner. Image translations involved subpixel shifts which resampled the images on new slightly modified coordinate systems. Image noise addition added randomized Gaussian noise based on the noise levels of the original slice. Volume alteration grew or shrank the mask based on the Euclidean distance transform and the percentage of volume added or subtracted. Lastly, contour randomization combined superpixel segmentation of the underlying image with a probabilistic selection of those superpixels based on their overlap with the mask to produce altered contours (**Figure 2**).

Each of the altered images then had its texture and topological features evaluated for the range of individual perturbations. For each category of perturbation and each feature, the intraclass correlation coefficient (ICC) was calculated to determine the variability or robustness of that feature to the perturbation in question. After calculating the ICC, any feature that had an ICC of <0.75 for any of the perturbations was excluded from this round of modeling. With that filter in place, the same modeling procedure was followed to evaluate the predictive power of texture and topological features across the 10 instances.

Classification With K-top Scoring Pairs

As an additional analysis, the same texture, topological, and CNN features were used to train a model using the k-top scoring pairs algorithm (kTSP). The kTSP algorithm classifies samples by identifying k-pairs of features whose relative expressions/values are inverted between the categories, i.e., it tries to find pairs of genes A and B whose relative rankings are inverted in most samples of the two cases. This gives an easy to interpret decision rule and makes the classifier robust to data normalization procedures. Given that different measurement technologies have different dynamic ranges, classifiers based on relative rankings of features rather than their absolute values are highly valuable for integrating and comparing across multiple sources of data.

TABLE 2 | Test set statistics for kTSP algorithm.

	AUROC	STD of AUROC
Texture only kTSP	0.659	0.099
TDA only kTSP	0.686	0.083
CNN only kTSP	0.718	0.111

Darker blue indicates improved AUROC.

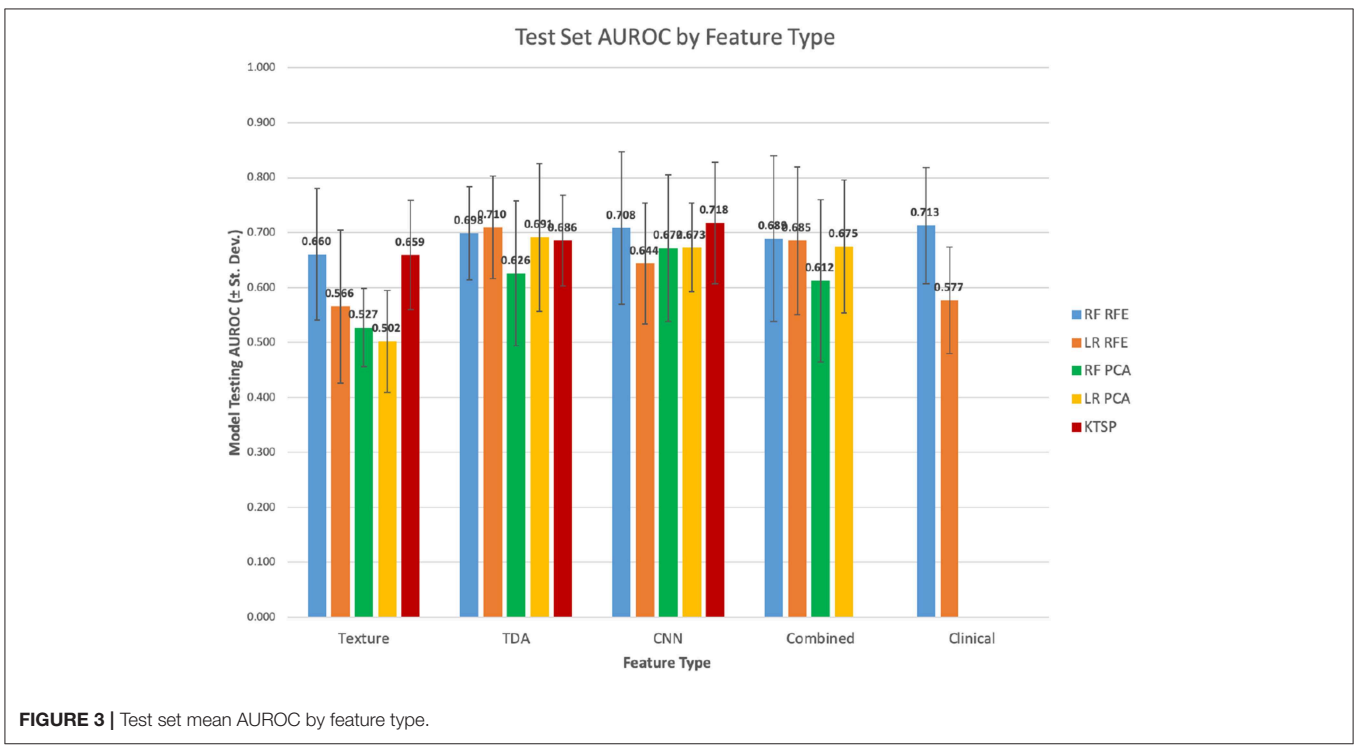


FIGURE 3 | Test set mean AUROC by feature type.

The extracted CNN, textural, and topological features were used to train a kTSP classifier for predicting patient 1p/19q codeletion status using the switchbox R package (Afsari et al., 2015). Since kTSP is a greedy algorithm, we retained only features that were measured to be significantly differential between the two classes (Wilcox test $p < 0.1$ after BH correction). We then split the data into training and test sets (70:30 split) and estimated classifier performance by measuring training and test set roc values. Since the codeletion cases were heavily resampled, we grouped features from the same patient together while doing the train/test split so as to ensure that training and testing cases are really independent. By repeating this procedure for a total of 5,000 times and building classifiers with k allowed to range between 3 and 15 pairs of features, we estimated the 95% highest posterior density intervals for train and test AUC values for classifiers built from the three datasets.

RESULTS

Texture features were evaluated across the 10 independent train/test splits to measure their predictive power (**Table 1**; **Figure 3**). PCA-based feature reduction on texture features did very poorly on the test set with an average AUROC across the 10 train/test splits of 0.502 with linear regression (LR) and 0.527 for random forest (RF). RFE achieved test set AUROC values of 0.660 and 0.566 for LR and the RF models, respectively. However, the standard deviation of AUROC across the different splits was quite high (0.120, 0.139), suggesting that with a small dataset, the models' performance can be somewhat unstable.

Features from topological data analysis were also evaluated across the 10-independent training/testing splits (**Table 1**). In this case, most of the analyses performed relatively similarly in terms of AUROC, ranging from 0.626 to 0.710 for these different models with topological features. Again, the standard deviation of AUROC across the different training/testing splits was relatively broad (0.085–0.135), though slightly lower than that of the texture features. Texture and TDA features overall had relatively similar performance, with a slight edge to TDA features, though well-within the variability of these statistics.

When modeled using random forests or logistic regression, the CNN feature set had similar predictive performance to topological features (**Table 1**). The AUROCs of these models fell between 0.644 and 0.708. It also performed similarly with the k-top scoring pairs (kTSP, **Table 2**) approach when compared to the random forest (RF) or logistic regression (LR) with an average AUROC of 0.718. Combining the three feature types neither improved or decreased performance, suggesting that they were not measuring vastly different types of information.

Overall, RFE somewhat outperformed PCA as a feature selection tool, although the scale of the difference depended on the feature set. Logistic regression had similar results to random forest classification in most cases, although there were some exceptions.

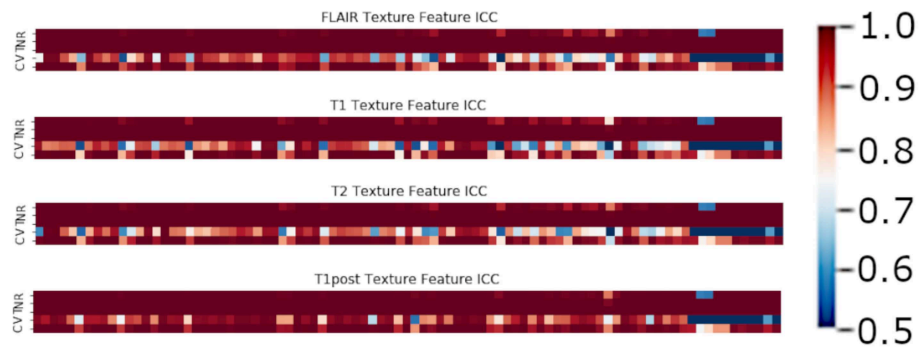


FIGURE 4 | Mean ICC of Texture features (R, Rotation; N, Noise addition; T, Translation; V, Volume alteration; C, Contour alteration). Volume based perturbations had the largest effect on the robustness of texture features, followed by contour alteration. There was a range of ICC values for the different features.

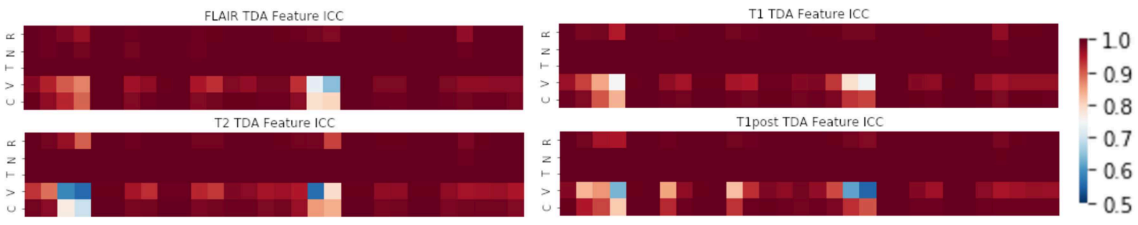


FIGURE 5 | Mean ICC of TDA features (R, Rotation; N, Noise addition; T, Translation; V, Volume alteration; C, Contour alteration). Volume based perturbations had the largest effect on ICC for topological features. Polynomial features 3 and 4 were the least robust to perturbation, while other TDA features were relatively stable.

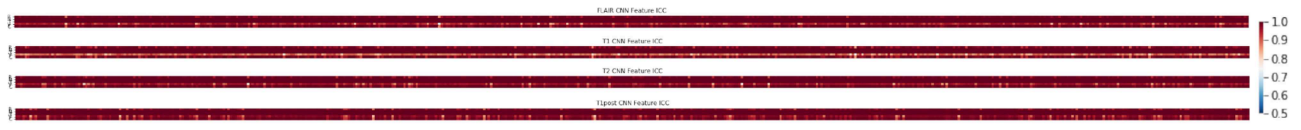


FIGURE 6 | Mean ICC of CNN features (R, Rotation; N, Noise addition; T, Translation; V, Volume alteration; C, Contour alteration). CNN based features were broadly stable to perturbations, though still most affected by volume changes.

In terms of feature robustness, topological features had much better ICC after perturbation than did the texture-based features. Of the 356 texture features, an average of ∼117 features (32.8%) had an ICC of <0.75 on the perturbations and were excluded from this round of modeling (**Figure 4**). Of the 120 topological features, an average of ∼10 (8.1%) had an ICC of <0.75, as such most of the features were included in the next round of modeling (**Figure 5**). Only an average of ∼3 of the CNN features (0.15%) were excluded at the 0.75 ICC cutoff (**Figure 6**).

The perturbation types which had the lowest average ICC were volume perturbation and contour alteration. Noise addition and translation had little impact on the ICC values for texture and TDA features. Volume alteration, and contour alteration both affect the segmentation mask of the tumor without an impact on the underlying image. This does, however, affect the region investigated by topological and texture features. Notably, when looking at stability in texture features by the class of feature, shape-based features performed poorly under volume-based alterations and were affected by rotation more than the other classes (**Figure 4**). Overall, GLCM-based measures were the most stable of the texture features as a class under these perturbations (**Figure 7**). Among the TDA features, polynomial features 3 and 4 were the least robust to perturbation, suggesting higher order polynomial features are less stable than lower order features (**Figure 5**).

Features that had a low ICC were excluded and the models were retrained on the reduced feature set. Then the predictive power of these models was measured on the testing set. Overall, when excluding non-robust features from modeling, the performance of the models dropped slightly in terms of AUROC, although most had relatively similar power (**Table 3**; **Figure 8**).

Increasing the ICC cutoff would increase the number of features excluded from the analysis. Thus, this effect was further studied for each type of image perturbation (**Figure 9**). Texture features are broadly susceptible to contour and volume alterations. A subset of texture features was susceptible to rotation effects as well, although very few features were affected by the noise or translation perturbations. CNN features had a relatively narrow range of ICC values, and TDA features were broadly stable, though a subset of TDA features were less robust.

DISCUSSION

In this study, topological data features performed as well as or better than texture features in predicting 1p/19q codeletion status. However, model performance varied across the different training and testing splits of the data, as evidenced by the standard deviation of model performance. CNN-based features also had similar performance to topological features with random forest and logistic regression, but they performed notably better with kTSP as the modeling algorithm.

One concern, however, is the relatively small sample size of 143 patients, of whom only 13 had the 1p/19q codeletion. This may be a large factor in the uncertainty in the prediction estimates. Oversampling the 1p/19q codeletion alleviates the class imbalance somewhat, but raises some concerns about overfitting, especially in models like random forest. Finding additional MRI studies with confirmed 1p/19q codeletion would improve the generalizability of any models derived from this data.

The kTSP algorithm is more often used in gene expression array data but can be applied just as easily to other large-scale datasets. By finding pairs of features that have different relative orderings in the two sets, kTSP is less dependent on the absolute magnitude of change than are the other methods. It also benefits from having a large number of features to search that have positive and negative associations with the target classification. As the CNN features are not human-designed features, and there is a larger set of CNN features with more variability in direction, kTSP seems to take better advantage of these features than features like TDA or texture.

Traditional radiomics features based on gray levels, such as GLCMs can be dependent on the number and boundaries of gray level bins. Volume and contour-based alterations affect the set of pixels under investigation, which could heavily influence the resulting texture matrices. Topological barcodes have been found to be mostly stable under image-based perturbations of the data, as have the CNN-based features from this pre-trained model.

While other groups have also used radiomic features or neural networks to predict 1p/19q codeletion, this paper seeks to compare multiple potential approaches (Han et al., 2018; Lu et al., 2018; Zhou et al., 2019). Other papers have trained neural networks to predict 1p/19q codeletion, whereas this study only used a pre-trained neural network on the dataset (Akkus et al., 2017; Chang et al., 2018b). One weakness of this approach was that the testing AUROCs of the models in this study were not as high as some that have been reported in other studies. However, this study was also able to evaluate the robustness of these features through image perturbation. Additionally, the models in this study incorporated topological features based on persistent homology, which had better performance than radiomic features and were more stable to perturbation.

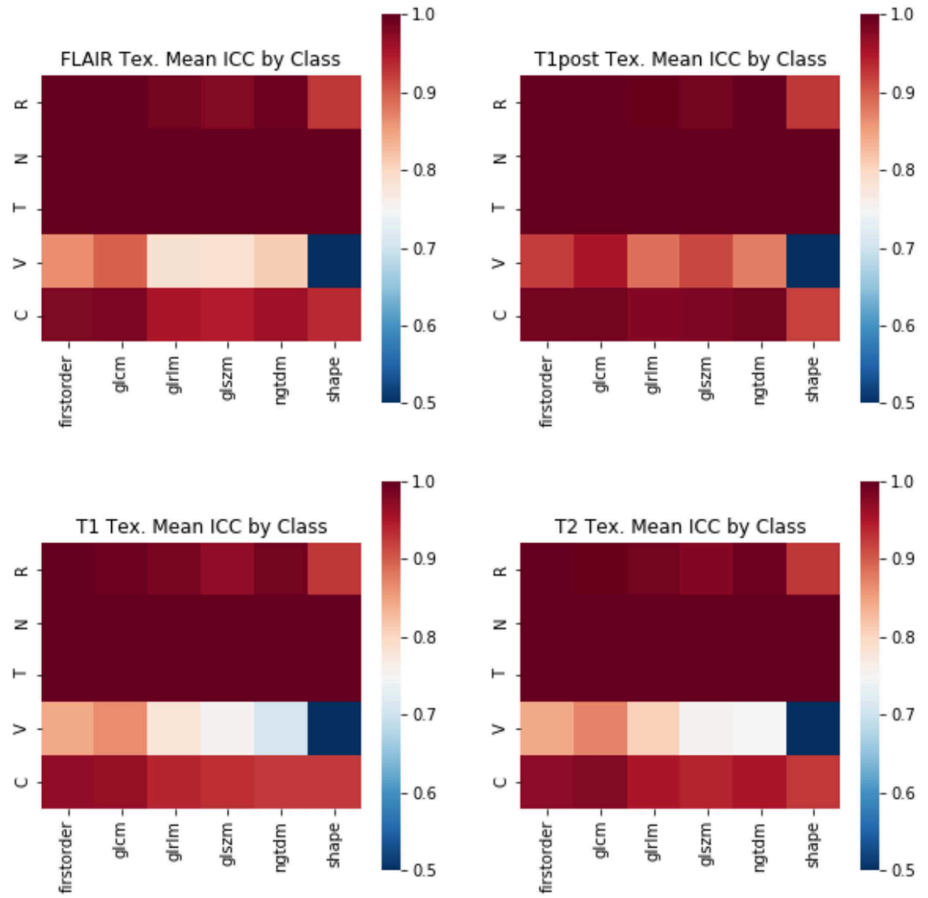


FIGURE 7 | Mean ICC of Texture features by feature class. (R, Rotation; N, Noise addition; T, Translation; V, Volume alteration; C, Contour alteration). Changes in volume had the largest effect on the stability of radiomic features. The least stable class of features were the shape-based features, whereas GLCM and first-order features were more stable.

TABLE 3 | Test set statistics, after exclusion of unstable features.

	AUROC	STD of AUROC	Sensitivity	Specificity	Accuracy
Texture only RF RFE	0.637	0.084	0.758	0.569	0.661
Texture only LR RFE	0.563	0.120	0.775	0.512	0.644
Texture only RF PCA	0.505	0.097	0.552	0.625	0.577
Texture only LR PCA	0.501	0.062	0.532	0.625	0.568
TDA only RF RFE	0.660	0.090	0.685	0.635	0.652
TDA only LR RFE	0.659	0.126	0.635	0.712	0.662
TDA only RF PCA	0.614	0.090	0.767	0.569	0.649
TDA only LR PCA	0.649	0.140	0.655	0.613	0.627
CNN only RF RFE	0.691	0.146	0.870	0.567	0.721
CNN only LR RFE	0.668	0.118	0.867	0.510	0.692
CNN only RF PCA	0.681	0.121	0.725	0.644	0.679
CNN only LR PCA	0.674	0.081	0.847	0.531	0.691
Combined RF RFE	0.681	0.146	0.860	0.552	0.707
Combined LR RFE	0.660	0.117	0.830	0.540	0.687
Combined RF PCA	0.650	0.163	0.760	0.619	0.686
Combined LR PCA	0.684	0.111	0.835	0.569	0.703

Darker blue indicates improved AUROC.

Clinical value is more difficult to assess than statistical significance, as it is dependent on the prognostic value of the biomarker, the current standard of care, and the predictive power of the model. 1p/19q codeletion is typically evaluated through genetic testing of a tissue sample, whereas the benefit of a radiogenomic approach is to evaluate the imaging markers of a tumor without biopsy or resection. However, as many glioma patients receive a biopsy for diagnostic purposes, a radiogenomic model would have to be exceptionally predictive to warrant replacement of this procedure. This study aims more to understand the types of features radiogenomic approaches are detecting, and how robust they are in different conditions rather than to replace the test.

FUTURE DIRECTIONS

While this study used the image perturbation parameter space of the Zwanenberg paper, it would be worthwhile to tune the tested space of parameters further. The level of noise is based on wavelet estimation, but by visual inspection is not apparent until the noise level is increased by 1–2 orders of

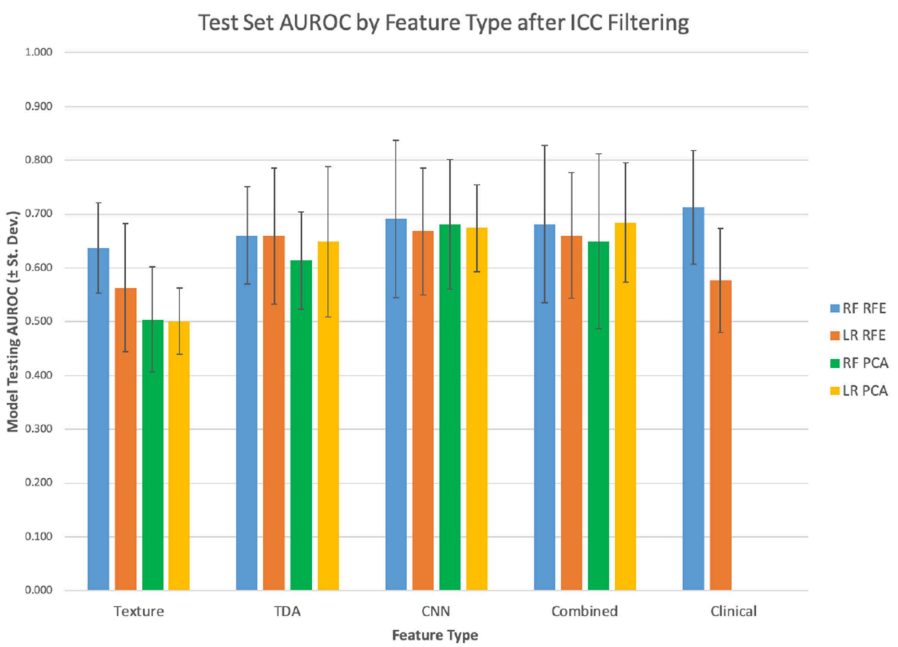


FIGURE 8 | Filtered features test set mean AUROC.

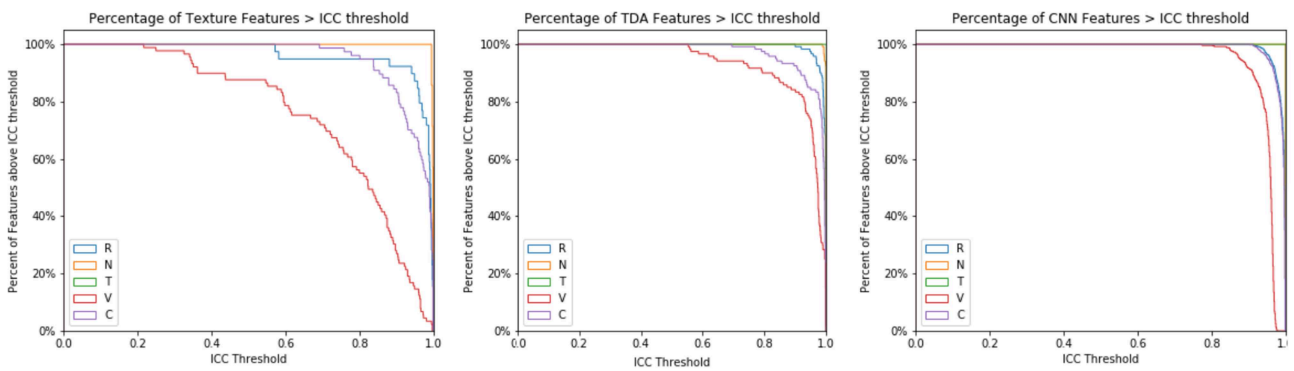


FIGURE 9 | ICC Threshold effect by perturbation type. The ICC cutoff for features is varied from 0 to 1, and the percentage of features that survive each threshold is recorded by perturbation type.

magnitude. Additional levels of noise could be investigated, as could the types of noise, such as changing the noise to a Rician distribution or adding the noise to k-space rather than the image domain. However, as these perturbations take each measurement and multiply it out by orders of magnitude, the computational demands can add up quickly. Thus, there is a tradeoff between perturbation complexity, the size of the parameter space, and the certainty of the resulting robustness measure.

Further investigation of the robustness of these measures could be done by simulating scans from the underlying physics, using a Bloch equation simulator (Ford et al., 2018). This would allow for measuring the effect of variable image collection parameters such as TE, TR, and field strength. Understanding these effects would help to account for concerns about variability in the underlying MRI protocols. Unfortunately, these simulations are primarily of normal brain images, so may not fully reflect the interaction between tumor tissue alteration and image feature robustness.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2017/data.html.

ETHICS STATEMENT

Human subjects data in the form of medical imaging data, and genetic information was collected from The Cancer Imaging Archive (TCIA) and The Cancer Genome Atlas (TCGA). All data in these National Cancer Institute (NCI) databases are anonymized and, therefore, individual institutional IRB approval is not required for this retrospective review, although it should be noted that all data were originally submitted to TCGA and TCIA by the contributing institutions under an IRB-approved protocol. TCGA LGG tumors were included in this study if they had a corresponding diagnostic MRI study in TCIA.

AUTHOR CONTRIBUTIONS

DK: initial development of modeling pipeline, data analysis and data normalization, and editing of manuscript. NW: reworked pipeline, data analysis, added image perturbation, and drafted manuscript. VR: development of kTSP section of code/data analysis, writing, and editing. DR: data analysis and manuscript editing. JL: kTSP data analysis and manuscript editing. AP: design of study and manuscript editing. RW: design of study and major revisions to manuscript. GR: design and conception of study. AR: design and oversight of study, provided direction for modeling, image perturbation and kTSP analysis, and manuscript editing.

FUNDING

AR, DK, and NW were supported by startup funds, Department of Radiology and MCubed grants from the University of Michigan, NCI 5R37CA214955-01A1 and a Research Scholar Grant from the American Cancer Society (RSG-16-005-01).

REFERENCES

Adcock, A., Carlsson, E., and Carlsson, G. (2013). The ring of algebraic functions on persistence bar codes. *arXiv[Preprint].arXiv:13040530*.

Adcock, A., Rubin, D., and Carlsson, G. (2014). Classification of hepatic lesions using the matching metric. *Comput. Vis. Image Underst.* 121, 36–42. doi: 10.1016/j.cviu.2013.10.014

Afsari, B., Fertig, E. J., Geman, D., and Marchionni, L. (2015). Switchbox: an R package for k–top scoring pairs classifier development. *Bioinformatics* 31, 273–274. doi: 10.1093/bioinformatics/btu622

Akkus, Z., Ali, I., Sedlár, J., Agrawal, J. P., Parney, I. F., Giannini, C., et al. (2017). Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. *J. Digit. Imaging* 30, 469–476. doi: 10.1007/s10278-017-9984-3

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). *Segmentation Labels for the Pre-operative Scans of the TCGA-LGG Collection*. The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017c). *Segmentation Labels for the Pre-Operative Scans of the TCGA-GBM Collection*. The Cancer Imaging Archive. Available online at: https://wiki.cancerimagingarchive.net/x/KoZyAQ (accessed April 5, 2019).

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv[Preprint].arXiv:181102629*.

Bauer, S., Fejes, T., and Reyes, M. (2012). A skull-stripping filter for ITK. *Insight J.* Available online at: http://hdl.handle.net/10380/3353

Bogowicz, M., Riesterer, O., Bundschuh, R. A., Veit-Haibach, P., Hüllner, M., Studer, G., et al. (2016). Stability of radiomic features in CT perfusion maps. *Phys. Med. Biol.* 61, 8736–8749. doi: 10.1088/1361-6560/61/24/8736

Boots-Sprenger, S. H. E., Sijben, A., Rijntjes, J., Tops, B. B. J., Idema, A. J., Rivera, A. L., et al. (2013). Significance of complete 1p/19q co-deletion, *IDH1* mutation and *MGMT* promoter methylation in gliomas: use with caution. *Mod. Pathol.* 26, 922–9. doi: 10.1038/modpathol.2012.166

Cairncross, G., Wang, M., Shaw, E., Jenkins, R., Brachman, D., Buckner, J., et al. (2013). Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402. *J. Clin. Oncol.* 31, 337–343. doi: 10.1200/JCO.2012.43.2674

Chang, K., Bai, H. X., Zhou, H., Su, C., Bi, W. L., Agbodza, E., et al. (2018a). Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin. Cancer Res.* 24, 1073–1081. doi: 10.1158/1078-0432.CCR-17-2236

Chang, P., Grinband, J., Weinberg, B. D., Bardis, M., Khy, M., Cadena, G., et al. (2018b). Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *Am. J. Neuroradiol.* 39, 1201–1207. doi: 10.3174/ajnr.A5667

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging.* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

Eckel-Passow, J. E., Lachance, D. H., Molinaro, A. M., Walsh, K. M., Decker, P. A., Sicotte, H., et al. (2015). Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *N. Engl. J. Med.* 372, 2499–2508. doi: 10.1056/NEJMoa1407279

Ford, J., Dogan, N., Young, L., and Yang, F. (2018). Quantitative radiomics: impact of pulse sequence parameter selection on MRI-based textural features of the brain. *Contrast. Media Mol. Imaging* 2018:1729071. doi: 10.1155/2018/1729071

Fuller, C. E., and Perry, A. (2005). Molecular diagnostics in central nervous system tumors. *Adv. Anat. Pathol.* 12:180. doi: 10.1097/01.pap.0000175117.47918.f7

Giansiracusa, N., Giansiracusa, R., and Moon, C. (2017). Persistent homology machine learning for fingerprint classification. *arXiv[Preprint].arXiv:171109158*.

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2015). Radiomics: images are more than pictures, they are data. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Han, Y., Xie, Z., Zang, Y., Zhang, S., Gu, D., Zhou, M., et al. (2018). Non-invasive genotype prediction of chromosome 1p/19q co-deletion by development and validation of an MRI-based radiomics signature in lower-grade gliomas. *J. Neurooncol.* 140, 297–306. doi: 10.1007/s11060-018-2953-y

Lu, C.-F., Hsu, F.-T., Hsieh, K. L.-C., Kao, Y.-C. J., Cheng, S.-J., Hsu, J. B.-K., et al. (2018). Machine learning–based radiomics for molecular subtyping of gliomas. *Clin. Cancer Res.* 24, 4429–4436. doi: 10.1158/1078-0432.CCR-17-3445

Maria, C. (2015). "Persistent cohomology," in *GUDHI User and Reference Manual*. (GUDHI Editorial Board). Available online at: http://gudhi.gforge.inria.fr/doc/latest/group__persistent__cohomology.html (accessed April 5, 2019).

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging.* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Van M den, B., Brandes, A. A., Taphoorn, M. J., Kros, J. M., Kouwenhoven, M. C., Delattre, J. Y., et al. (2013). Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 31, 344–350. doi: 10.1200/JCO.2012.43.2229

van Timmeren, J. E., Leijenaar, R. T. H., van Elmpt, W., Wang, J., Zhang, Z., Dekker, A., et al. (2016). Test–retest data for radiomics feature stability analysis: generalizable or study-specific? *Tomography* 2, 361–365. doi: 10.18383/j.tom.2016.00208

Zhou, H., Chang, K., Bai, H. X., Xiao, B., Su, C., Bi, W. L., et al. (2019). Machine learning reveals multimodal MRI patterns predictive of isocitrate dehydrogenase and 1p/19q status in diffuse low- and high-grade gliomas. *J. Neurooncol.* 142, 299–307. doi: 10.1007/s11060-019-03096-0

Zwanenburg, A., Leger, S., Agolli, L., Pilz, K., Troost, E. G. C., Richter, C., et al. (2019). Assessing robustness of radiomic features by image perturbation. *Sci. Rep.* 9:614. doi: 10.1038/s41598-018-36938-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 13 August 2019
doi: 10.3389/fncom.2019.00056

Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation

Guotai Wang[1,2]*, *Wenqi Li*[2,3], *Sébastien Ourselin*[2] and *Tom Vercauteren*[2]

[1] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China, [2] School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, [3] NVIDIA, Cambridge, United Kingdom

Automatic segmentation of brain tumors from medical images is important for clinical assessment and treatment planning of brain tumors. Recent years have seen an increasing use of convolutional neural networks (CNNs) for this task, but most of them use either 2D networks with relatively low memory requirement while ignoring 3D context, or 3D networks exploiting 3D features while with large memory consumption. In addition, existing methods rarely provide uncertainty information associated with the segmentation result. We propose a cascade of CNNs to segment brain tumors with hierarchical subregions from multi-modal Magnetic Resonance images (MRI), and introduce a 2.5D network that is a trade-off between memory consumption, model complexity and receptive field. In addition, we employ test-time augmentation to achieve improved segmentation accuracy, which also provides voxel-wise and structure-wise uncertainty information of the segmentation result. Experiments with BraTS 2017 dataset showed that our cascaded framework with 2.5D CNNs was one of the top performing methods (second-rank) for the BraTS challenge. We also validated our method with BraTS 2018 dataset and found that test-time augmentation improves brain tumor segmentation accuracy and that the resulting uncertainty information can indicate potential mis-segmentations and help to improve segmentation accuracy.

Keywords: brain tumor segmentation, deep learning, uncertainty, data augmentation, convolutional neural network

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Mauricio Reyes,
University of Bern, Switzerland
Alle Meije Wink,
VU University Medical Center,
Netherlands
Siddhesh Pravin Thakur,
University of Pennsylvania,
United States

***Correspondence:**
Guotai Wang
guotai.wang@uestc.edu.cn

Received: *24 April 2019*
Accepted: *30 July 2019*
Published: *13 August 2019*

Citation:
Wang G, Li W, Ourselin S and Vercauteren T (2019) Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation. Front. Comput. Neurosci. 13:56. doi: 10.3389/fncom.2019.00056

1. INTRODUCTION

In adults, gliomas are the most common primary brain tumors. They begin in the brain's glial cells and are typically categorized into different grades: High-Grade Gliomas (HGG) grow rapidly and are more malignant, while Low-Grade Gliomas (LGG) are slower growing tumors with a better patient prognosis (Louis et al., 2016). Magnetic Resonance Imaging (MRI) of brain tumors is critical for progression evaluation, treatment planning and assessment of this disease. Different sequences of MRI can be used for brain tumor imaging, such as T1-weighted, T2-weighted, contrast enhanced T1-weighted (T1ce), and Fluid Attenuation Inversion Recovery (FLAIR) images. T2 and FLAIR images mostly highlight the whole tumor region (including infiltrative edema), and T1 and T1ce images give a better contrast for the tumor core region (not including infiltrative edema) (Menze et al., 2015). Therefore, these different sequences providing complementary information can be combined for the analysis of different subregions of brain tumors.

Segmenting brain tumors and subregions automatically from multi-modal MRI is important for reproducible and accurate measurement of the tumors, and this can assist better diagnosis, treatment planning and evaluation (Menze et al., 2015; Bakas et al., 2017b). However, it remains difficult for automatic methods to accurately segment brain tumors from multi-modal MRI. This is due to the fact that the images often have ambiguous boundaries between normal tissues and brain tumors. In addition, though prior information of shape and position has been used for segmentation of anatomical structures such as the liver (Wang et al., 2015) and the heart (Grosgeorge et al., 2013), the shape, size and position of brain tumors have considerable variations across different patients. This makes it difficult to use a prior shape and position for robust segmentation of brain tumors. Recently, deep learning methods with Convolutional Neural Networks (CNNs) have become the state-of-the-art approaches for brain tumor segmentation (Bakas et al., 2018). Compared with traditional supervised learning methods such as decision trees (Zikic et al., 2012) and support vector machines (Lee et al., 2005), CNNs can learn the most useful features automatically, without the need for manual design and selection of features.

A key problem for CNN-based segmentation is to design a suitable network structure and training strategy. Using a 2D CNN in a slice-by-slice manner has a relatively low memory requirement (Havaei et al., 2016), but the network ignores 3D information, which will ultimately limit the performance of the segmentation. Using 3D CNNs can better exploit 3D features, but requires a large amount of memory, which may limit the input patch size, depth or feature numbers of the CNNs (Kamnitsas et al., 2017b). As a trade-off, 2.5D CNNs can take advantage of inter-slice features compared with 2D CNNs and have a lower memory requirement than their 3D counterparts. In addition, whole tumor, tumor core and enhancing tumor core follow a hierarchical structure. Using the segmentation of whole tumor (tumor core) to guide the segmentation of tumor core (enhancing tumor core) can help to reduce false positives. Therefore, in this work, we propose a framework consisting of a cascade of 2.5D networks for brain tumor segmentation from multi-modal 3D MRI that achieves a trade-off between memory consumption, model complexity and receptive field.

For medical images, uncertainty information of segmentation results is important for clinical decisions as it can help to understand the reliability of the segmentations (Shi et al., 2011) and identify challenging cases necessitating expert review (Jungo et al., 2018). For example, for brain tumor images, the low contrast between surrounding tissues and the segmentation target leads voxels around the boundary to be labeled with less confidence. The uncertainty information of these voxels can indicate regions that have potentially been mis-segmented, and therefore can be employed to guide interactions of human to refine the segmentation results (Wang et al., 2018b). In addition, compared with datasets for natural image recognition (Russakovsky et al., 2015), datasets for CNN-based medical image segmentation methods are relatively small, which tends to result in more uncertain predictions in the segmentation outputs, and can lead to structure-wise uncertainty for downstream tasks, such as measuring the volume of tumor regions. Therefore, this work also aims at providing voxel-wise and structure-wise uncertainty information for CNN-based brain tumor segmentation. Unlike model-based (*epistemic*) uncertainty obtained by test-time dropout (Gal and Ghahramani, 2016; Jungo et al., 2017, 2018), we investigate image-based (*aleatoric*) uncertainty obtained by test-time augmentation that has previously been mainly used for improving segmentation accuracy (Matsunaga et al., 2017; Radosavovic et al., 2018).

This paper is a combination and an extension of our previous works on brain tumor segmentation (Wang et al., 2017, 2018a), where we proposed a cascade of CNNs for sequential segmentation of brain tumor and the subregions from multi-modal MRI, which decomposes the complex task of multi-class segmentation into three simpler binary segmentation tasks. We also proposed 2.5D network structures with anisotropic convolution for the segmentation task as a result of trade-off between memory consumption, model complexity and receptive field. In this paper, we extend them in two aspects. First, we use test-time augmentation to obtain uncertainty estimation of the segmentation results, and additionally propose an uncertainty-aware conditional random field (CRF) for post-processing. The results show that uncertainty estimation not only helps to identify potential mis-segmentations but also can be used to improve segmentation performance. Both voxel-level and structure-level uncertainty are analyzed in this paper. Second, we implement more ablation studies to demonstrate the effectiveness of our segmentation pipeline.

2. RELATED WORKS

2.1. Brain Tumor Segmentation From MRI

Existing brain tumor segmentation methods include generative and discriminative approaches. By incorporating domain-specific prior knowledge, generative approaches usually have good generalization to unseen images, as they directly model probabilistic distributions of anatomical structures and textural appearances of healthy tissues and the tumor (Menze et al., 2010). However, it is challenging to precisely model probabilistic distributions of brain tumors. In contrast, discriminative approaches extract features from images and associate the features with the tissue classes using discriminative classifiers. They often require a supervised learning set-up where images and voxel-wise class labels are needed for training. Classical methods of this category include decision trees (Zikic et al., 2012) and support vector machines (Lee et al., 2005).

Recently, CNNs as a type of discriminative approach have achieved promising results on multi-modal brain tumor segmentation tasks. Havaei et al. (2016) combined local and global 2D features extracted by a CNN for brain tumor segmentation. Although it outperformed the conventional discriminative methods, the 2D CNN only uses 2D features without considering the volumetric context. To incorporate 3D features, applying the 2D networks in axial, sagittal and coronal

views and fusing their results has been proposed (McKinley et al., 2016; Li and Shen, 2017; Hu et al., 2018). However, the features employed by such a method are from cross-planes rather than entire 3D space.

DeepMedic (Kamnitsas et al., 2017b) used a 3D CNN to exploit multi-scale volumetric features and further encoded spatial information with a fully connected Conditional Random Field (CRF). It achieved better segmentation performance than using 2D CNNs but has a relatively low inference efficiency due to the multi-scale image patch-based analysis. Isensee et al. (2018) applied 3D U-Net to brain tumor segmentation with a carefully designed training process. Myronenko (2018) used an encoder-decoder architecture for 3D brain tumor segmentation and the network contained an additional branch of variational auto-encoder to reconstruct the input image for regularization. To obtain robust brain tumor segmentation resutls, Kamnitsas et al. (2017a) proposed an ensemble of multiple CNNs including 3D Fully Convolutional Networks (FCN) (Long et al., 2015), DeepMedic (Kamnitsas et al., 2017b), and 3D U-Net (Ronneberger et al., 2015; Abdulkadir et al., 2016). The ensemble model is relatively robust to the choice of hyper-parameters of each individual CNN and reduces the risk of overfitting. However, it is computationally intensive to run a set of models for both training and inference (Malmi et al., 2015; Pereira et al., 2017; Xu et al., 2018).

2.2. Uncertainty Estimation for CNNs

Uncertainty information can come from either the CNN models or the input images. For model-based (*epistemic*) uncertainty, exact Bayesian modeling is mathematically grounded but often computationally expensive and hard to implement. Alternatively, Gal and Ghahramani (2016) cast test-time dropout as a Bayesian approximation to estimate a CNN's model uncertainty. Zhu and Zabaras (2018) estimated uncertainty of a CNN's parameters using approximated Bayesian inference via stochastic variational gradient descent. Other approximation methods include Monte Carlo batch normalization (Teye et al., 2018), Markov chain Monte Carlo (Neal, 2012) and variational Bayesian (Louizos and Welling, 2016). Lakshminarayanan et al. (2017) proposed a simple and scalable method using ensembles of models for uncertainty estimation. For test image-based (*aleatoric*) uncertainty, Ayhan and Berens (2018) found that test-time augmentation was an effective and efficient method for exploring the locality of a test sample in *aleatoric* uncertainty estimation, but its application to medical image segmentation has not been investigated. Kendall and Gal (2017) proposed a unified Bayesian framework that combines *aleatoric* and *epistemic* uncertainty estimations for deep learning models. In the context of brain tumor segmentation, Eaton-Rosen et al. (2018) and Jungo et al. (2018) used test-time dropout to estimate the uncertainty. Wang et al. (2019a) analyzed a combination of *epistemic* and *aleatoric* uncertainties for whole tumor segmentation, but the uncertainty information of other structures (tumor core and enhancing tumor core) was not investigated.

3. METHODS

3.1. Segmentation Pipeline and Network Structure

3.1.1. Triple Cascaded Framework

Malmi et al. (2015) and Pereira et al. (2017) used a cascade of two stages to segment brain tumors where the whole tumor was segmented in the first stage and then all substructures were segmented in the second stage. To better take advantage of the hierarchical property of brain tumor structures, in our preliminary study (Wang et al., 2017), we proposed a cascade of three CNNs to hierarchically and sequentially segment the whole brain tumor, tumor core and enhancing tumor core, which is followed by some more recent works (Ma and Yang, 2018; Xu et al., 2018). As shown in **Figure 1**, we use three networks (WNet, TNet, and ENet) to segment these structures, respectively. First, the whole tumor is segmented by WNet. Then the input multi-modal image is cropped according to the bounding box of the segmented whole tumor. Second, TNet segments the tumor core from the cropped image region, and the input image is further cropped based on the bounding box of the segmented tumor core. Finally, the enhancing tumor core is segmented by ENet from the second cropped region. We use the segmentation result of whole tumor (tumor core) as a crisp mask for the result of tumor core (enhancing tumor core), which leads to anatomical constraints for the final segmentation.

3.1.2. Anisotropic Convolutional Neural Networks

To achieve a trade-off between memory consumption, model complexity and receptive field for 3D brain tumor segmentation, we propose anisotropic 2.5D CNNs with a large intra-slice receptive field and a relatively small inter-slice receptive field. These CNNs take a stack of slices as input. The receptive field of WNet and TNet is $217 \times 217 \times 9$, and that of ENet is $113 \times 113 \times 9$. **Figure 2** shows structures of these proposed CNNs. Note that in previous works (McKinley et al., 2016; Li and Shen, 2017), fusing 2D networks in three orthogonal views was referred to as a 2.5D network, where each of the single-view networks only captures 2D features. In our method, we also use multi-view fusion, but the network in each view is a 2.5D network that captures anisotropic 3D features.

The anisotropic receptive field of our CNNs is achieved by decomposing a typical 3D $3 \times 3 \times 3$ convolution kernel into an intra-slice convolution kernel and an inter-slice convolution kernel, with kernel size of $3 \times 3 \times 1$ and $1 \times 1 \times 3$, respectively. We use four inter-slice convolution layers and 20 intra-slice convolution layers in the backbone of our CNNs, and set the output channel number of these convolution layers to a fixed number C_0. To facilitate the training process, batch normalization is used after each convolution, as shown in green and blue blocks in **Figure 2**. He et al. (2015) found that Parametric Rectified Linear Units (PReLU) outperforms traditional rectified units, therefore we use PReLU as our activation function. Two 2D downsampling layers are used to reduce the resolution of feature maps of WNet and TNet while avoiding large loss of segmentation details. ENet shares the same

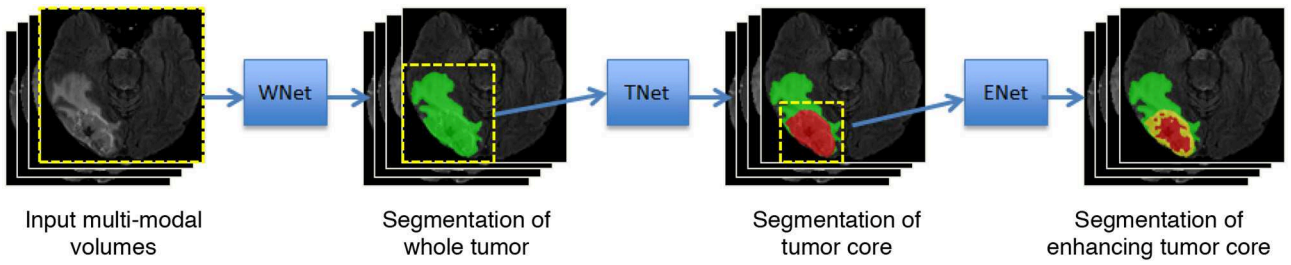


FIGURE 1 | Our proposed framework with triple cascaded CNNs for multi-modal brain tumor segmentation. We use three CNNs to hierarchically and sequentially segment whole tumor, tumor core and enhancing tumor core, and the CNNs are referred to as WNet, TNet, and ENet, respectively.

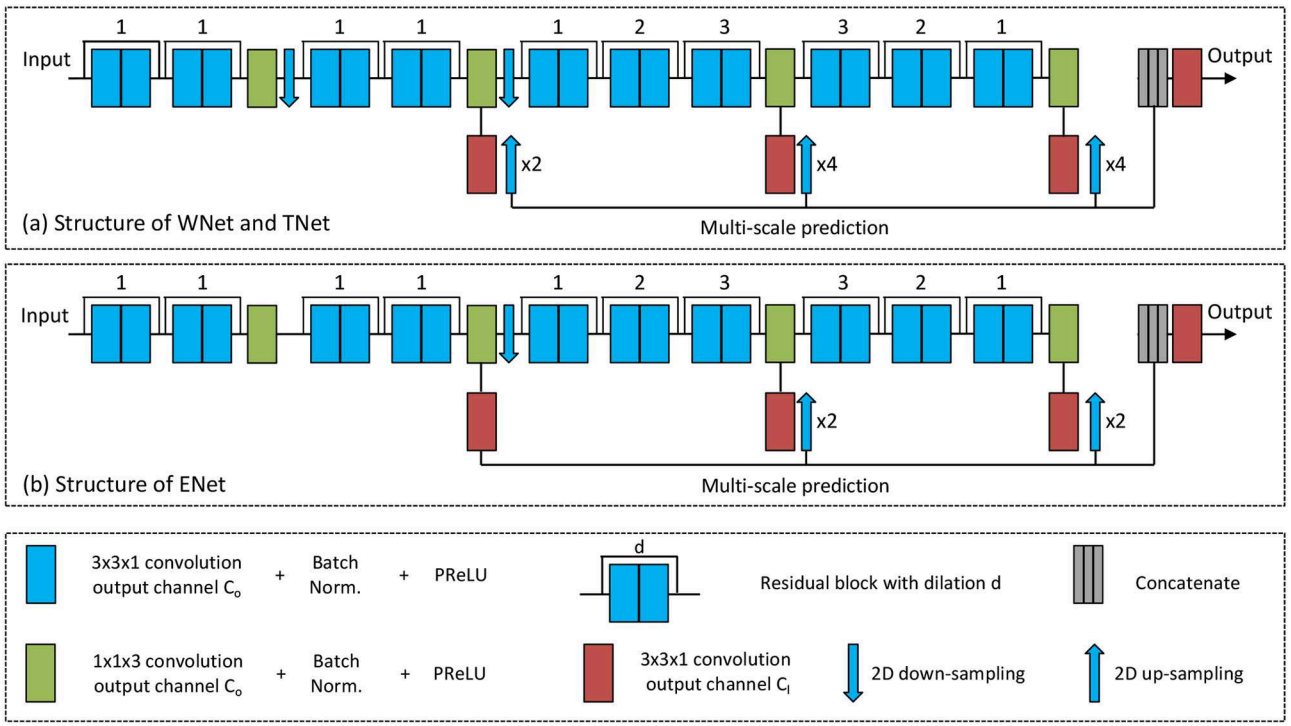


FIGURE 2 | The proposed anisotropic CNNs with residual connection, dilated convolution, and multi-scale prediction. Only one downsampling layer is used in ENet as its input size is smaller.

structure with WNet and TNet except that it uses only one downsampling layer, as the input size of ENet is smaller.

As shown in **Figure 2**, intra-slice convolution layers are grouped into 10 blocks, and each block includes two intra-slice convolution layers. To speed the convergence of training, we use residual connections (He et al., 2016) by adding the output of each block directly to its input. We also employ dilated convolution to increase the intra-slice receptive field. The dilation parameter is shown on the top of each residual block in **Figure 2**. In addition, each CNN uses multi-scale prediction for deep supervision. To get multiple intermediate predictions, three prediction layers with $3\times3\times1$ convolution are used at different depths of the CNNs, as depicted by red boxes in **Figure 2**. These intermediate predictions are upsampled to the resolution of the input and concatenated. An additional prediction layer with $3\times3\times1$ convolution is used to obtain the final score map from the concatenated intermediate predictions. The output channel number of these prediction layers is denoted as C_l, and is set to 2 in this paper.

3.1.3. Multi-view Fusion

The above anisotropic CNNs have a small through-plane receptive field, and therefore have a limited ability to make use of 3D contextual information. To overcome this problem, we use multi-view fusion where all WNet, TNet, and ENet are trained in three orthogonal (axial, sagittal, and coronal) views, respectively. At test time, for each network structure, we use the corresponding versions of trained models to obtain the segmentation results in these three views, respectively, and average their softmax outputs to obtain a single fused result.

3.2. Augmentation for Training and Testing

Considering the image acquisition process, one underlying anatomy can be observed with different conditions, such as various spatial transformations and intensity noise. Therefore, an acquired image can be seen as only one of many possible observations of the target. Directly applying CNNs to the single observed image may lead the result to be biased toward the specific transformation and noise in the given observation. To address this problem, we predict the segmentation result by considering different spatial transformations and intensity noise for a test image.

Let β denote spatial transformation parameters and e represent intensity noise, respectively. Though all images in the BraTS datasets are aligned to a standard orientation, we use rotation, flipping and scaling to augment the variation of local features. Therefore, we represent β as a composition of r, f_l and s, where r denotes the rotation angle along each spatial axis in 3D, f_l is a random binary value representing flipping along each 3D axis or not, and s denotes a scaling factor. We consider some prior distributions of these parameters: $r \sim U(0, 2\pi)$, $f_l \sim Bern(0.5)$, and $s \sim U(0.8, 1.2)$. In addition, we assume that the intensity noise follows a prior distribution of $e \sim N(0, 0.05)$ according to Wang et al. (2019a).

To obtain augmented images, we use Monte Carlo simulation to randomly sample β and e from the above prior distributions N times, and each time we use the sampled parameters to generate a transformed image. The augmentation process is used at both training and testing stage for a given network. For test-time augmentation, the Monte Carlo simulation leads to N transformed versions of the same input image, and they are fed into the CNN for inference. We combine the N predicted results via majority voting to obtain the final prediction of each structure.

3.3. Uncertainty Estimation of Segmentation Results

3.3.1. Voxel-Wise Uncertainty

In our method, the use of test-time augmentation provides multiple prediction results of the same input image with different spatial transformations and intensity changes. The disagreement between these predictions naturally gives an uncertainty estimation of the segmentation. Therefore, we use test-time augmentation to obtain not only segmentation results but also the associated image-based (*aleatoric*) uncertainty. Differently from Wang et al. (2019a), we provide uncertainty estimation not only for the whole tumor, but also for the substructures (tumor core and enhancing tumor core).

To obtain voxel-wise uncertainty estimation, we measure the diversity of the N different predictions for a given voxel in the test image. Let X and Y represent the input image and the output segmentation, respectively, and let Y^i represent the i-th voxel's predicted label. Typically, the uncertainty of Y^i can be estimated by the entropy and variance of the distribution of Y^i, rather than averaged probability map resulting from N Monte Carlo samples that cannot reflect the diversity information. For multi-class segmentation of BraTS, the variance of discrete class label for a voxel is not sufficiently representative. Therefore, we use entropy of Y^i to estimate the voxel-wise uncertainty, which is desired for image segmentation tasks. Assume a set of N discrete values (i.e., labels) for Y^i is denoted as $\mathcal{Y}^i = \{y_1^i, y_2^i, \ldots, y_N^i\}$, then we can approximate the entropy of the distribution of Y^i by:

$$H(Y^i|X) \approx -\sum_{m=1}^{M} \hat{p}_m^i \ln(\hat{p}_m^i) \tag{1}$$

where $\hat{p}_m^i$ is the frequency of the m-th unique value in $\mathcal{Y}^i$. When $\mathcal{Y}^i$ is obtained by test-time augmentation with Monte Carlo simulation described in section 3.2, Equation (1) represents voxel-wise *aleatoric* uncertainty.

3.3.2. Structure-Wise Uncertainty

The above Monte Carlo simulation obtains N segmentation results for a given structure in a test image. For the i-th simulation, let v_i denote the volume of the segmented structure, then the set of volumes of the N segmentations is denoted as $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$. Assume that the mean value and standard deviation of $\mathcal{V}$ is $\mu_{\mathcal{V}}$ and $\sigma_{\mathcal{V}}$, respectively. Then the structure-wise uncertainty is estimated as the volume variation coefficient (VVC):

$$VVC = \frac{\sigma_{\mathcal{V}}}{\mu_{\mathcal{V}}} \tag{2}$$

In this paper, $\mathcal{V}$ is obtained by test-time augmentation, leading Equation (2) to represent structure-wise *aleatoric* uncertainty.

4. EXPERIMENTS AND RESULTS

4.1. Data and Implementation Details

We validated our methods with the BraTS 2017[1] and BraTS 2018[2] (Menze et al., 2015; Bakas et al., 2017a,b) datasets. The two datasets share the same set of training images from 285 patients, including 75 cases of LGG and 210 cases of HGG. The validation sets of BraTS 2017 and BraTS 2018 contain images from 46 and 66 patients with brain tumors respectively. The testing sets of BraTS 2017 and BraTS 2018 contain images from 146 and 191 patients with brain tumors, respectively. The grades of brain tumors in the validation and testing sets are unknown. Each patient was scanned with FLAIR, T1ce, T1, and T2. The original images were acquired across different views and the resolution was anisotropic. All the images had been re-sampled to an isotropic 1.0 mm × 1.0 mm × 1.0 mm resolution and skull-striped by the organizers. In addition, the four modalities of the same patient had been co-registered. As the BraTS organizers provided ground truth only for the training set, we randomly selected 20% from the training set as our local validation set during training.

Our 2.5D CNNs were implemented in Tensorflow[3] (Abadi et al., 2016) using NiftyNet[4][5] (Gibson et al., 2018). We used

[1] http://www.med.upenn.edu/sbia/brats2017.html
[2] http://www.med.upenn.edu/sbia/brats2018.html
[3] https://www.tensorflow.org
[4] http://niftynet.io
[5] https://github.com/NifTK/NiftyNet/tree/dev/demos/BRATS17

an NVIDIA TITAN X GPU with 12 GB memory, Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014) and Dice loss function (Milletari et al., 2016; Fidon et al., 2017a) for training, with batch size 5, weight decay 10^{-7}, initial learning rate 10^{-3}, and iteration number 30k. The training patch size was 144 × 144 × 19 for WNet, and 96 × 96 × 19 and 64 × 64 × 19 for TNet and ENet, respectively. We normalized each image by the intensity mean and standard deviation, and set the channel number C_o of intermediate convolution layers to 32 and class number C_l to 2. We trained all WNet, TNet and ENet for axial, sagittal and coronal views separately as our networks had a relatively small number of parameters. Therefore, each network had three different sets of parameters. At test time, the predictions in these three views were averaged. We applied training-time and test-time augmentation to BraTS 2018 dataset according to 3.2, and the Monte Carlo simulation number N was set to 20. We uploaded our segmentation results of the validation and testing datasets to the publicly available evaluation server of BraTS 2017 and BraTS 2018, and the server gave quantitative evaluation results in terms of Dice score and Hausdorff distance.

4.1.1. Results of BraTS 2017 Dataset

4.1.1.1. Qualitative results

We first validated our proposed segmentation framework with BraTS 2017 dataset, and test-time augmentation was not used for this experiment. We compared our proposed cascade of anisotropic networks with multi-view fusion with two variants: (1) cascade of 3D isotropic networks that captures 3D features directly, where we remove all 1×1×3 convolutions in WNet, TNet and ENet, and replace 3×3×1 convolutions and 2D down-sampling (up-sampling) with 3×3×3 convolutions and 3D donw-sampling (up-sampling), respectively, and this variant is referred to as isotropic 3D networks; (2) cascade of our anisotropic networks but without multi-view fusion, where the networks are only implemented in axial view, and this variant is referred to as anisotropic 2.5D networks.

Figure 3 shows two examples for HGG and LGG segmentation from our local validation set that is a subset of BraTS 2017/2018 training set. We only show the FLAIR images in the inputs of CNNs for simplicity of visualization. Edema, non-enhancing tumor core and enhancing tumor core are visualized in green, red and yellow, respectively. The results of isotropic 3D networks and anisotropic 2.5D networks are shown in the second and third rows, respectively. In the case of HGG shown in **Figure 3A**, isotropic 3D networks obtain some mis-segmentations of the edema, and anisotropic 2.5D networks result in some noise in the edema and enhancing tumor core regions. In contrast, the proposed method leads to more accurate segmentation results. **Figure 3B** shows a case of LGG that does not contain enhancing tumor core. The segmentation results of whole tumor are similar for the three methods. However, the proposed method outperforms isotropic 3D networks and anisotropic 2.5D networks in the tumor core region.

4.1.1.2. Quantitative evaluation

Quantitative evaluation results with the BraTS 2017 validation set are shown in **Table 1**. The average Dice scores achieved by our method for enhancing tumor core, whole tumor and tumor core are 0.786, 0.905 and 0.838, respectively, which outperforms isotropic 3D networks and anisotropic 2.5D networks. We also compared our method with Kamnitsas et al. (2017a) that uses an ensemble of multiple CNNs for segmentation, and Isensee et al. (2017) that combines 3D U-Net with residual connection and deep supervision. **Table 1** shows that our method outperforms the others on the BraTS 2017 validation set. The quantitative evaluation results of our method on BraTS 2017 testing set are shown in **Table 2**. According to the BraTS 2017 organizers[6], our method won the second place of the BraTS 2017 segmentation task, while Kamnitsas et al. (2017a) and Isensee et al. (2017) ranked in the first and third place, respectively.

4.1.2. Results of BraTS 2018 Dataset

We then applied our proposed segmentation framework to BraTS 2018 dataset. To validate the effect of test-time augmentation (TTA), we compared three network configurations as underpinning CNNs: (1) 3D UNet (Abdulkadir et al., 2016) reimplemented by NiftyNet, (2) our cascaded networks where the whole tumor, tumor core and enhancing tumor core were segmented by WNet, TNet, and ENet, respectively, and (3) adapting WNet for multi-class segmentation without using a cascade of binary predictions, where we changed the output channel number for prediction layers to 4. We refer to this variant as multi-class WNet and also use multi-view fusion for it. The 3D U-Net and multi-class WNet were trained in the same way as our cascaded networks.

4.1.2.1. Qualitative results

Figure 4 shows two examples from the BraTS 2018 validation set. In each subfigure, the input images (FLAIR, T1, T1ce, and T2) are shown in the first row and the segmentation results of different networks with and without TTA are presented in the second row. In **Figure 4A**, the result of 3D UNet without TTA contains some false positives in the edema and non-enhancing tumor core regions. In contrast, the result of 3D UNet + TTA is more spatially consistent. The result obtained by multi-class WNet without TTA also contains some noise for the segmented non-enhancing tumor core, and multi-class WNet + TTA obtains a smoother segmentation. It can also be observed that our cascaded CNNs + TTA performs better on the tumor core than the counterpart without TTA. In **Figure 4B**, 3D UNet seems to obtain an under-segmentation in the central part of the tumor core, and 3D UNet + TTA overcomes this under-segmentation. Multi-class WNet without TTA seems to have an over segmentation for the non-enhancing tumor core region, and the counterpart with TTA achieves a higher accuracy in contrast. For our cascaded CNNs, TTA also helps to improve the spatial consistency of the segmentation result in this case.

4.1.2.2. Quantitative evaluation

Table 3 shows the quantitative evaluation results of different approaches on the validation set of BraTS 2018. Dice scores achieved by 3D UNet without TTA for enhancing tumor core,

[6]https://www.med.upenn.edu/sbia/brats2017/rankings.html

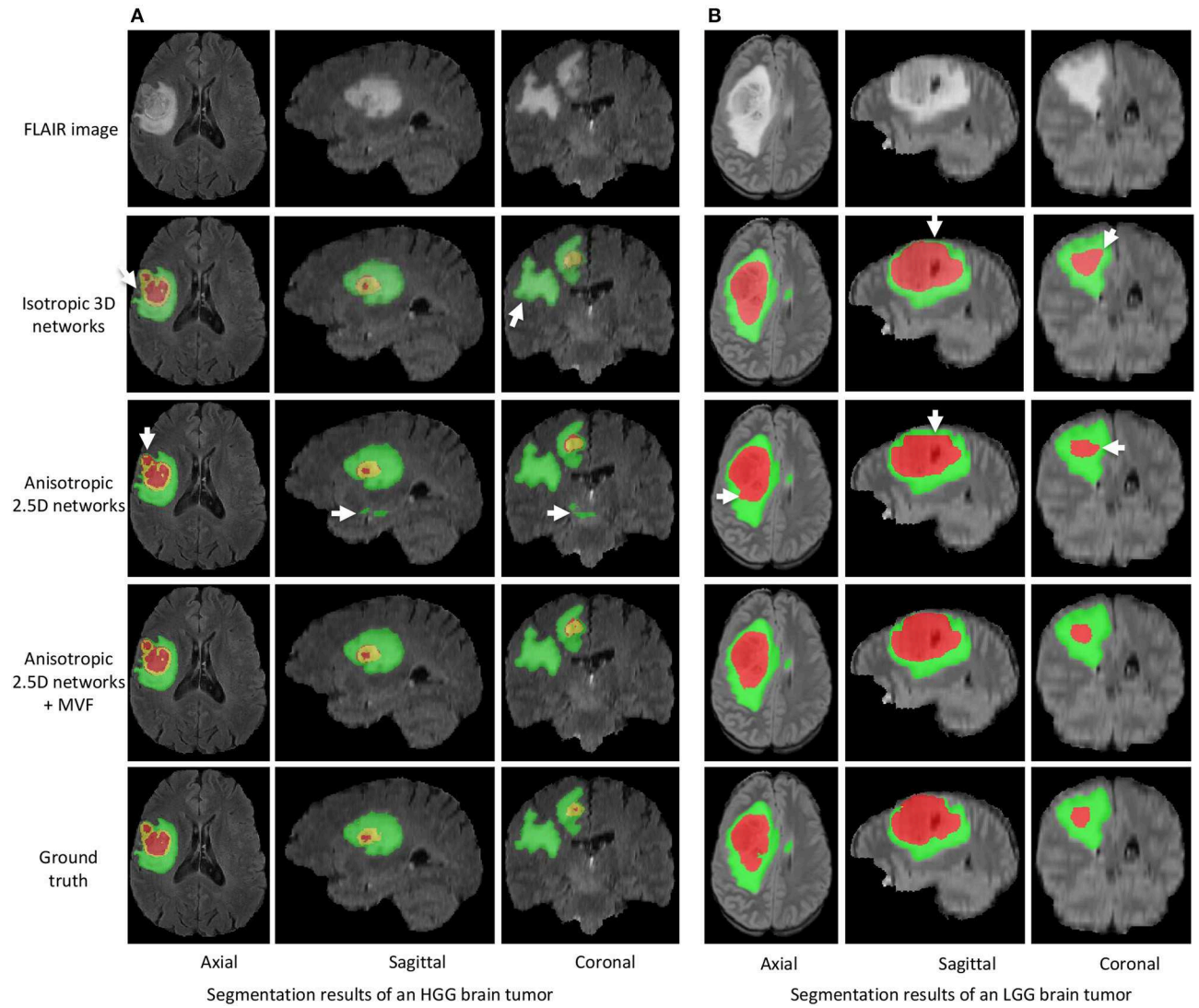


FIGURE 3 | Segmentation results of an HGG brain tumor **(A)** and an LGG brain tumor **(B)** from our local validation set, which is part of BraTS 2017/2018 training set. Edema, non-enhancing tumor core and enhancing tumor core are visualized in green, red, and yellow, respectively. White arrows highlight some mis-segmentations.

TABLE 1 | Dice and Hausdorff distance of our method on validation set of BraTS 2017 (mean ± std).

	Dice			Hausdorff (mm)		
	ET	**WT**	**TC**	**ET**	**WT**	**TC**
Isotropic 3D networks	0.772 ± 0.268	0.885 ± 0.105	0.805 ± 0.196	3.78 ± 5.32	6.73 ± 9.19	7.75 ± 9.98
Anisotropic 2.5D networks	0.741 ± 0.264	0.890 ± 0.076	0.826 ± 0.157	5.32 ± 7.20	12.46 ± 21.47	9.66 ± 14.21
Our method	**0.786 ± 0.233**	**0.905 ± 0.066**	**0.838 ± 0.158**	**3.28 ± 3.88**	**3.89 ± 2.79**	**6.48 ± 8.26**
Kamnitsas et al., 2017a	0.738	0.901	0.797	4.50	4.23	6.56
Isensee et al., 2017	0.732	0.896	0.797	4.55	6.97	9.48

MVF, multi-view fusion; ET, enhancing tumor core; WT, whole tumor; TC, tumor core. Our method: cascaded framework with anisotropic 2.5D CNNs and MVF. Bold value shows the best performance.

whole tumor and tumor core are 0.734, 0.864 and 0.766, respectively. Combining TTA with 3D UNet achieved a better performance, leading to Dice scores of 0.754, 0.873, and 0.783 for these structures, respectively. Applying test-time augmentation to multi-class WNet and the cascaded networks also leads to an improvement of segmentation accuracy. We also compared our method with Myronenko (2018) and Isensee et al. (2018) that ranked the first and second of BraTS 2018 segmentation

TABLE 2 | Dice and Hausdorff distance of our method on testing set of BraTS 2017 (mean ± std).

	Dice			Hausdorff (mm)		
	ET	**WT**	**TC**	**ET**	**WT**	**TC**
Our method	**0.783 ± 0.222**	0.874 ± 0.132	0.775 ± 0.270	**15.90 ± 67.86**	6.55 ± 10.69	27.05 ± 84.43
Kamnitsas et al., 2017a	0.729	**0.886**	**0.785**	36.0	**5.01**	**23.10**
Isensee et al., 2017	0.647 ± 0.326	0.858 ± 0.161	0.775 ± 0.269	–	–	–

ET, enhancing tumor core; WT, whole tumor; TC, tumor core. Bold value shows the best performance.

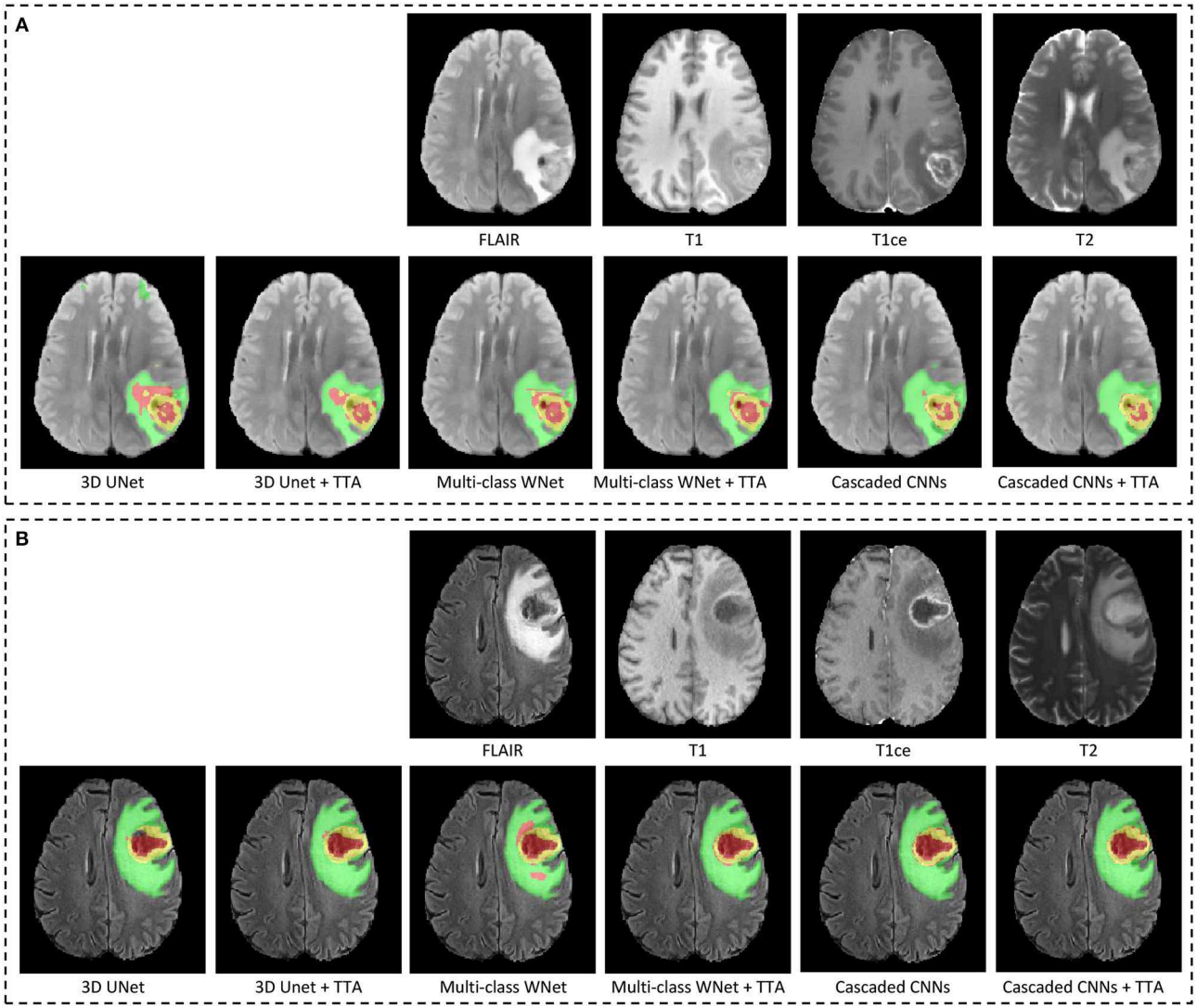


FIGURE 4 | Examples of test-time augmentation (TTA) combined with different CNNs for brain tumor segmentation. The images are from BraTS 2018 validation set, of which ground truth are not provided by the organizer. In each subfigure, the first row shows the input image of the same patient in four modalities, and the second row shows segmentation results. Edema, non-enhancing tumor core and enhancing tumor core are visualized in green, red, and yellow, respectively. **(A,B)** Show images of two different patients.

challenge, respectively[7]. Myronenko (2018) used an ensemble of 10 models, and we list the result of a single model and that of model ensemble reported by Myronenko (2018). Isensee et al. (2018) trained a 3D U-Net with additional datasets for the segmentation task. It can be observed that our method performs closely to these two compared methods on BraTS 2018 validation

[7]https://www.med.upenn.edu/sbia/brats2018/rankings.html

TABLE 3 | Dice and Hausdorff distance of different methods on validation set of BraTS 2018 (mean ± std).

	Dice			Hausdorff (mm)		
	ET	**WT**	**TC**	**ET**	**WT**	**TC**
3D UNet	0.734 ± 0.284	0.864 ± 0.146	0.766 ± 0.230	9.37 ± 22.95	12.00 ± 21.22	10.37 ± 13.47
3D UNet + TTA	0.754 ± 0.263	0.873 ± 0.125	0.783 ± 0.168	4.53 ± 9.60	5.90 ± 6.80	8.03 ± 10.31
Multi-class WNet	0.757 ± 0.257	0.890 ± 0.089	0.725 ± 0.245	4.24 ± 7.97	4.99 ± 6.53	12.13 ± 13.41
Multi-class WNet + TTA	0.771 ± 0.242	0.896 ± 0.071	0.730 ± 0.255	4.44 ± 8.20	4.92 ± 6.42	11.13 ± 13.46
Cascaded networks	0.792 ± 0.233	0.903 ± 0.057	0.854 ± 0.142	3.34 ± 4.15	5.38 ± 9.31	6.61 ± 8.55
Cascaded networks + TTA	0.797 ± 0.229	0.902 ± 0.056	0.858 ± 0.139	3.13 ± 3.78	6.18 ± 9.53	6.37 ± 8.19
Cascaded networks + TTA + CRF0	0.803 ± 0.228	0.905 ± 0.056	0.862 ± 0.136	3.09 ± 3.75	5.97 ± 8.22	6.25 ± 7.87
Cascaded networks + TTA + CRF1	0.807 ± 0.225	0.908 ± 0.054	**0.869 ± 0.126**	3.01 ± 3.69	5.86 ± 8.16	**6.09 ± 7.74**
Myronenko, 2018 (single model)	0.815	0.904	0.860	3.80	**4.48**	8.28
Myronenko, 2018 (ensemble)	**0.823**	**0.910**	0.867	3.93	4.52	6.85
Isensee et al., 2018	0.810	0.908	0.854	**2.54**	4.97	7.04

ET, enhancing tumor core; WT, whole tumor; TC, tumor core; TTA, test-time augmentation. CRF0: naive conditional random field for post-processing. CRF1: our uncertainty-aware conditional random field. Bold value shows the best performance.

TABLE 4 | Dice and Hausdorff evaluation of our cascaded CNNs with test-time augmentation (TTA) on testing set of BraTS 2018 (mean ± std).

	Dice			Hausdorff (mm)		
	ET	**WT**	**TC**	**ET**	**WT**	**TC**
Cascaded networks + TTA	0.747 ± 0.259	0.878 ± 0.119	0.796 ± 0.250	4.16 ± 7.07	5.97 ± 8.56	6.71 ± 10.27
Myronenko, 2018	0.766 ± 0.256	0.884 ± 0.118	0.815 ± 0.250	3.77 ± 8.61	5.90 ± 10.01	4.81 ± 7.52
Isensee et al., 2018	0.779 ± 0.239	0.878 ± 0.129	0.806 ± 0.250	2.90 ± 3.85	6.03 ± 9.98	5.08 ± 8.09

ET, enhancing tumor core; WT, whole tumor; TC, tumor core. Myronenko (2018) used an ensemble of 10 models for the segmentation.

set. Quantitative evaluation results of our cascaded CNNs with TTA on BraTS 2018 testing set is presented in **Table 4**. The results are compared with those of Myronenko (2018) and Isensee et al. (2018). Note that Myronenko (2018) requires a large amount of GPU memory (32 GB) for training, and Isensee et al. (2018) trained the model with additional datasets. **Table 4** shows that the segmentation accuracy of our proposed framework is comparable with that of the other two counterparts.

4.1.2.3. Uncertainty estimation

Figure 5 presents a case from our local validation set of BraTS 2018, where **Figures 5C,D** show the results of our cascaded CNNs and the corresponding voxel-wise uncertainty obtained by TTA, respectively. It can be observed that most uncertain results concentrate on the border of the tumor's substructures and some regions that are potentially mis-segmented. The white arrow in **Figure 5C** highlights a region that has been mis-segmented by CNNs, and the corresponding region has high uncertainty values in **Figure 5D**. To investigate the usefulness of the uncertainty information for improving segmentation accuracy, we reset the foreground and background probability of voxels with uncertainty higher than a threshold value (i.e., 0.2) to 0.5, and then use a conditional random field (CRF) for post-processing. This method is referred to as uncertainty-aware CRF, and it is compared with a naive CRF that is applied to the probability output of CNNs directly. **Figures 5E,F** show that the uncertainty-aware CRF outperforms the naive CRF for post-processing. **Table 3** shows a quantitative comparison between these post-processing methods using and not using uncertainty information on validation set of BraTS 2018.

We also measured structure-wise uncertainty based on VVC defined in Equation (2) for BraTS 2018 validation set. **Figure 6** shows the relationship between structure-wise segmentation error in terms of 1-Dice and uncertainty in terms of VVC. The figure shows that for all the three structures of enhancing tumor core, whole tumor and tumor core, a higher VVC value tends to be linked with a higher segmentation error. This demonstrates that the structure-wise uncertainty based on our test-time augmentation is informative and it can indicate potential mis-segmentations.

5. DISCUSSION AND CONCLUSION

The proposed cascaded system is well-suited for hierarchical tumor region segmentation. Compared with using a single network for multi-class segmentation, its main advantages are: (1) The use of three binary segmentation networks decomposes the complex task of multi-class segmentation and allows for a simpler network for each sub-task. They reduce the risk

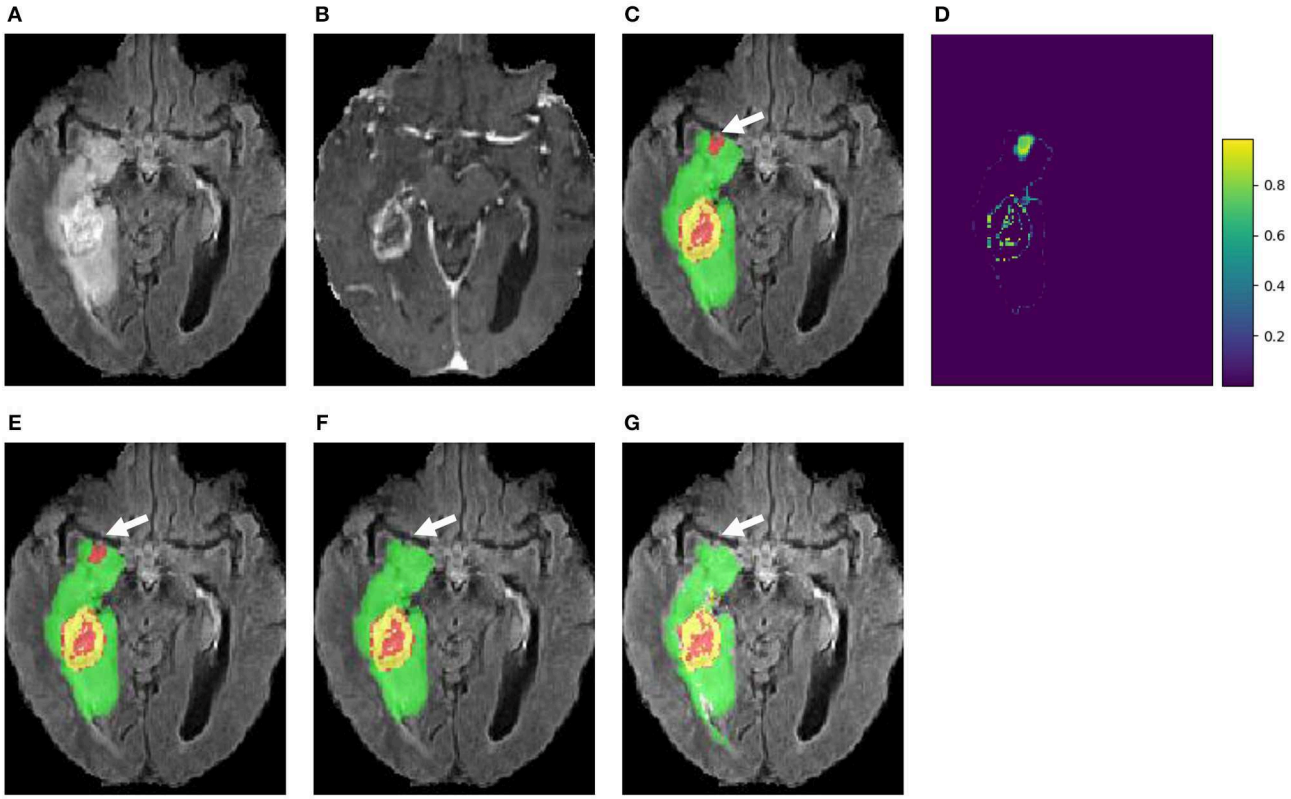


FIGURE 5 | An example of brain tumor segmentation result and the associated voxel-wise uncertainty estimation based on our cascaded CNNs with test-time augmentation (TTA). Taking the uncertainty information for post-processing by conditional random field (CRF) helps to correct the mis-segmented region, as shown in **(F)**. **(A)** FLAIR, **(B)** T1ce, **(C)** Initial segmentation, **(D)** Voxel-wise uncertainty, **(E)** Post-process with CRF, **(F)** Post-process with uncertainty-aware CRF, and **(G)** Ground truth.

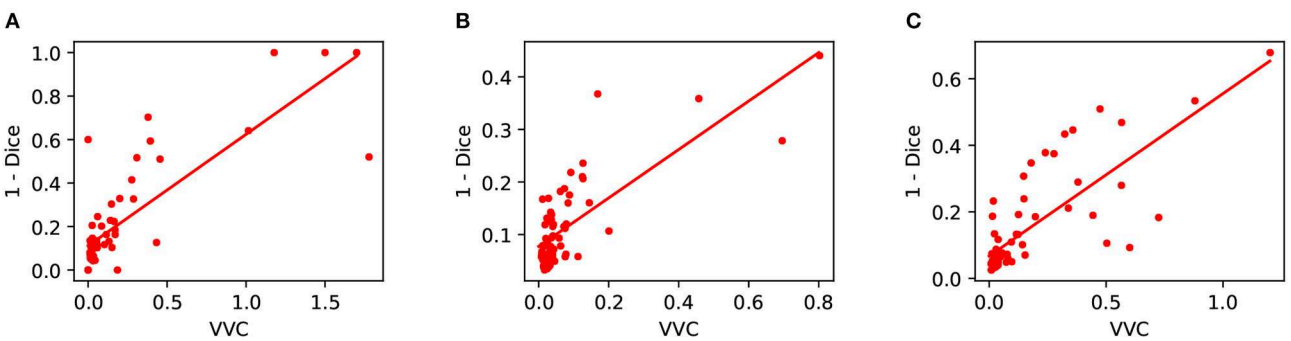


FIGURE 6 | Relationship between segmentation error (1-Dice) and structure-wise uncertainty in terms of volume variation coefficient (VVC) for BraTS 2018 validation set. **(A)** Enhancing core, **(B)** Whole tumor, and **(C)** Tumor core.

of over-fitting and are easier to train. (2) The cascade can effectively reduce the number of false positives because a subsequent network (e.g., TNet) only works on the image region selected by its precedent network (e.g., WNet). (3) The decomposition of the segmentation task also imposes strong spatial constraints which follows the anatomical structures of the brain tumor. It is also possible to model the hierarchical nature of the labels by adopting task-specific loss functions (e.g., Fidon et al., 2017a). However, Fidon et al. (2017a) did not use the hierarchical structural information as spatial constraints. Unlike most works that optimize the segmentation based on mutually exclusive edema, necrotic, and enhancing tumor core, our method optimizes the hierarchical whole tumor, tumor core and enhancing tumor core. This leads to the idea

of training networks on such loss criteria to simultaneously obtain these hierarchical structures in a single forward pass, as demonstrated by Myronenko (2018). For some clinical cases where the tumor does not have edema component, i.e., the region of whole tumor is the same as that of tumor core, our model may encounter some difficulties (e.g., false positives of edema) as all the training data in our experiments include edema region. However, as our WNet segments the edema region and tumor core region as a whole, the tumor core region in such cases will not be missed in the output of WNet. It is of interest to validate the proposed method on such cases in the future. In addition, in our cascaded segmentation framework, segmentation of whole tumor (tumor core) was used as a crisp mask for tumor core (enhancing tumor core), this may lead mis-segmentations in an early stage to cause mis-segmentations in a later stage. It would be of interest to investigate a better solution to combine the results obtained in different stages.

Compared with the single multi-class network approach using similar network structures, the training and inference of our proposed cascade require a longer time. In practice, we found that it is not a critical issue for automatic brain tumor segmentation. In fact, the inference of our method is more efficient than many competitive approaches such as DeepMedic (Kamnitsas et al., 2017b) and ScaleNet (Fidon et al., 2017b).

The multi-view fusion is an important component of the proposed system (as demonstrated in **Figure 3**). It is designed to combine the outputs from the lightweight and anisotropic networks applied in different views so that the 3D contextual information is fully utilized. To further incorporate different imaging resolutions in the multi-view fusion, it might be helpful to consider a weighted combination of the orthogonal views rather than a simple arithmetic mean (Mortazi et al., 2017).

From **Table 3** we find that the improvement obtained by TTA varies for different networks. For 3D UNet (Abdulkadir et al., 2016), the performance improvement is considerable, especially for the Hausdorff distance. For our cascaded networks, the improvement is relatively smaller but TTA is also effective to reduce the distance errors for enhancing tumor and tumor core. **Table 3** also shows that TTA reduces the standard deviation (improves the robustness) of the networks in most cases, especially for 3D UNet. For our cascaded networks, the standard deviations for enhancing tumor and tumor core are also smaller when TTA is used. Therefore, TTA can be seen as a robustness booster. In the proposed system, data augmentation only includes adding random intensity noise and spatial transformations such as rotation, flipping and scaling. It is also possible to adopt more complex transformations such as elastic deformations (Abdulkadir et al., 2016).

We have investigated the test image-based (*aleatoric*) uncertainty for brain tumor segmentation using test-time augmentation. We additionally show that the uncertainty information can be leveraged to improve the segmentation accuracy, as demonstrated in **Table 3** and **Figure 5**. The obtained uncertainty could be useful for downstream analysis such as uncertainty-aware volume measurement (Eaton-Rosen et al., 2018) and guiding user interactions (Wang et al., 2018b). Combining *epistemic* uncertainty based on test-time dropout or CNN ensembles (Kamnitsas et al., 2017a; Myronenko, 2018) and *aleatoric* uncertainty based on test-time augmentation is also an interesting future direction. It should be noticed that current methods for BraTS challenge heavily rely on voxel-wise annotations, which is difficult and time-consuming to collect for large datasets. In the future, it is of interest to learn from weakly or partially annotated brain tumor images in a larger dataset and improve generalizability of the CNNs. Some of the automatically segmented results can also be interactively refined to improve the robustness of brain tumor segmentation for clinic use (Wang et al., 2019b).

In conclusion, we have developed a novel system consisting of a cascade of 2.5D CNNs for brain tumor segmentation from multi-modal MRI, which decomposes the multi-class segmentation task into three sequential binary segmentation tasks. The 2.5D CNNs consider the balance between memory consumption, model complexity and recpetive field, and are combined with multi-view fusion for robust segmentation. We also studied the effect of combining test-time augmentatiofn with CNNs in the segmentation task and investigated the resulting *aleatoric* uncertainty estimation for the segmentation results. Experimental results based on BraTS 2017 dataset showed that our method was one of the top-performing methods. Experiments also showed that test-time augmentation led to an improvement of segmentation accuracy for different CNN structures and effectively obtained voxel-wise and structure-wise uncertainty estimation of the segmentation results that helps to improve segmentation accuracy.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2018/data.html.

AUTHOR CONTRIBUTIONS

GW, WL, and TV contributed conception and design of the study. GW and WL contributed implementation of the method. GW conducted the experiments and wrote the manuscript. All authors contributed to manuscript revision, proofreading, and approved the submitted version.

FUNDING

This work was supported by the Wellcome Trust [WT101957, WT97914, 203145/Z/16/Z, 203148/Z/16/Z], Engineering and Physical Sciences Research Council (EPSRC) [NS/A000027/1, NS/A000049/1, NS/A000050/1], hardware donated by NVIDIA. TV is supported by a Medtronic/Royal Academy of Engineering Research Chair [RCSRF1819/7/34].

ACKNOWLEDGMENTS

We would like to thank the NiftyNet team.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "TensorFlow: A system for large-scale machine learning," in *USENIX Symposium on Operating Systems Design and Implementation* (Savannah, GA), 265–284.

Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Athens), 424–432.

Ayhan, M. S., and Berens, P. (2018). "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," in *Medical Imaging with Deep Learning* (Amsterdam), 1–9.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Arch.* doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nat. Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Alessandro, C., et al. (2018). 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.* doi: 10.17863/CAM.38755

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., and Cardoso, M. J. (2018). "Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Granada), 691–699.

Fidon, L., Li, W., and Garcia-peraza herrera, L. C. (2017a). "Generalised Wasserstein Dice score for imbalanced multi-class segmentation using holistic convolutional networks," in *International MICCAI Brainlesion Workshop* (Quebec, QC), 64–76.

Fidon, L., Li, W., Garcia-Peraza-Herrera, L. C., Ekanayake, J., Kitchen, N., Ourselin, S., et al. (2017b). "Scalable multimodal8 convolutional networks for brain tumour segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Quebec, QC), 285–293.

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning* (New York, NY), 1050–1059.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., et al. (2018). NiftyNet: A deep-learning platform for medical imaging. *Comput. Methods Prog. Biomed.* 158, 113–122. doi: 10.1016/j.cmpb.2018.01.025

Grosgeorge, D., Petitjean, C., Dacher, J. N., and Ruan, S. (2013). Graph cut segmentation with a statistical shape model in cardiac MRI. *Comput. Vis. Image Underst.* 117, 1027–1035. doi: 10.1016/j.cviu.2013.01.014

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2016). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *ICCV* (Santiago), 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *CVPR* (Las Vegas, NV), 770–778.

Hu, Y., Liu, X., Wen, X., Niu, C., and Xia, Y. (2018). "Brain tumor segmentation on multimodal MR imaging using multi-level upsampling in decoder yan," in *International MICCAI Brainlesion Workshop* (Granada), 168–177.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2017). "Brain tumor segmentation8 and radiomics survival prediction: contribution to the BRATS 2017 challenge," in *International MICCAI Brainlesion Workshop* (Quebec, QC), 287–297.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Granada), 234–244.

Jungo, A., McKinley, R., Meier, R., Knecht, U., Vera, L., Perez-Beteta, J., et al. (2017). "Towards uncertainty-assisted brain tumor segmentation and survival prediction," in *International MICCAI Brainlesion Workshop* (Quebec, QC), 474–485.

Jungo, A., Meier, R., Ermis, E., Herrmann, E., and Reyes, M. (2018). Uncertainty-driven sanity check: application to postoperative brain tumor cavity segmentation. *arXiv [Preprint]. arXiv1806.03106.* Available online at: https://arxiv.org/abs/1806.03106

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2017a). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI Brainlesion Workshop* (Quebec, QC), 450–462.

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017b). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kendall, A., and Gal, Y. (2017). "What uncertainties do we need in Bayesian deep learning for computer vision?" in *NeurIPS* (Long Beach, CA), 5580–5590.

Kingma, D. P., and Ba, J. L. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]. arXiv:1412.6980.* Available online at: https://hdl.handle.net/11245/1.505367

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS* (Long Beach, CA), 6405–6416.

Lee, C.-H., Schmidt, M., and Murtha, A. (2005). "Segmenting brain tumors with conditional random fields and support vector machines," in *International Workshop on Computer Vision for Biomedical Image Applications* (Beijing), 469–478.

Li, Y., and Shen, L. (2017). "Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries," in *Deep Learning Based Multimodal Brain Tumor Diagnosis* (Quebec, QC), 149–158.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *CVPR* (Boston, MA), 3431–3440.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Louizos, C., and Welling, M. (2016). "Structured and efficient variational eeep learning with matrix gaussian posteriors," in *ICML* (New York, NY), 1708–1716.

Ma, J. and Yang, X. (2018). "Automatic brain tumor segmentation by exploring the multi-modality complementary information and cascaded 3D lightweight CNNs," in *International MICCAI Brainlesion Workshop* (Granada: Springer International Publishing), 25–36.

Malmi, E., Parambath, S., Peyrat, J.-M., Abinahed, J., and Chawla, S. (2015). "CaBS: A cascaded brain tumor segmentation approach," in *Proceeding MICCAI BRATS Challenge* (Munich), 42–47.

Matsunaga, K., Hamada, A., Minagawa, A., and Koga, H. (2017). Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv [Preprint]. arXiv1703.03108.* Available online at: https://arxiv.org/abs/1703.03108

McKinley, R., Wepfer, R., Gundersen, T., Wagner, F., Chan, A., Wiest, R., et al. (2016). "Nabla-net: A deep dag-like convolutional architecture for biomedical image segmentation," in *Int. MICCAI Brainlesion Work* (Athens), 119–128.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Menze, B. H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., and Golland, P. (2010). "A generative model for brain tumor segmentation in multi-modal images," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Beijing), 151–159.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *IC3DV* (Stanford, CA), 565–571.

Mortazi, A., Karim, R., Rhode, K., Burt, J., and Bagci, U. (2017). "CardiacNET: Segmentation of left atrium and proximal pulmonary veins from MRI using multi-view CNN," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Quebec, QC), 377–385.

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Worksho* (Granada). p. 349–356.

Neal, R. M. (2012). *Bayesian Learning for Neural Networks*. Toronto, ON: Springer Science & Business Media.

Pereira, S., Oliveira, A., Alves, V., and Silva, C. A. (2017). "On hierarchical brain tumor segmentation in MRI using fully convolutional neural networks: a preliminary study," in *IEEE 5th Portuguese Meeting on Bioengineering (ENBENG)* (Coimbra), 1–4.

Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., and He, K. (2018). "Data distillation: towards omni-supervised learning," in *CVPR* (Salt Lake City, UT), 4119–4128.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Munich), 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Shi, W., Zhuang, X., Wolz, R., Simon, D., Tung, K., Wang, H., et al. (2011). "A multi-image graph cut approach for cardiac image segmentation and uncertainty estimation," in *International Workshop on Statistical Atlases and Computational Models of the Heart* (Toronto, ON), 178–187.

Teye, M., Azizpour, H., and Smith, K. (2018). "Bayesian uncertainty estimation for batch normalized deep networks," in *International Conference on Machine Learning* (Stockholm).

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019a). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338, 34–45. doi: 10.1016/j.neucom.2019.01.103

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Quebec, QC), 178–190.

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2018a). "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in *International MICCAI Brainlesion Workshop, Vol. 10670* (Granada), 61–72.

Wang, G., Li, W., Zuluaga, M. A., Pratt, R., Patel, P. A., Aertsen, M., et al. (2018b). Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Trans. Med. Imaging* 37, 1562–1573. doi: 10.1109/TMI.2018.2791721

Wang, G., Zhang, S., Xie, H., Metaxas, D. N., and Gu, L. (2015). A homotopy-based sparse representation for fast and accurate shape prior modeling in liver surgical planning. *Med. Image Anal.* 19, 176–186. doi: 10.1016/j.media.2014.10.003

Wang, G., Zuluaga, M. A., Li, W., Pratt, R., Patel, P. A., Aertsen, M., et al. (2019b). DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1559–1572. doi: 10.1109/TPAMI.2018.2840695

Xu, Y., Gong, M., Fu, H., Tao, D., and Zhang, K. (2018). "Multi-scale masked 3-D U-net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Granada: Springer International Publishing), 222–233.

Zhu, Y., and Zabaras, N. (2018). Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *J. Comput. Phys.* 366, 415–447. doi: 10.1016/j.jcp.2018.04.018

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., et al. (2012). "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Nice), 369–376.

Conflict of Interest Statement: WL was employed by King's College London during most of the preparation of this work and was employed by company NVDIA for the final editing and proofreading of the manuscript. SO is a founder and shareholder of BrainMiner Ltd, UK.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

METHODS
published: 16 August 2019
doi: 10.3389/fnins.2019.00810

Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning

Li Sun[1], *Songtao Zhang*[1], *Hang Chen*[1] *and Lin Luo*[1,2]*

[1] School of Innovation and Entrepreneurship, Southern University of Science and Technology, Shenzhen, China, [2] College of Engineering, Peking University, Beijing, China

Gliomas are the most common primary brain malignancies. Accurate and robust tumor segmentation and prediction of patients' overall survival are important for diagnosis, treatment planning and risk factor identification. Here we present a deep learning-based framework for brain tumor segmentation and survival prediction in glioma, using multimodal MRI scans. For tumor segmentation, we use ensembles of three different 3D CNN architectures for robust performance through a majority rule. This approach can effectively reduce model bias and boost performance. For survival prediction, we extract 4,524 radiomic features from segmented tumor regions, then, a decision tree and cross validation are used to select potent features. Finally, a random forest model is trained to predict the overall survival of patients. The 2018 MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS), ranks our method at 2nd and 5th place out of 60+ participating teams for survival prediction tasks and segmentation tasks respectively, achieving a promising 61.0% accuracy on the classification of short-survivors, mid-survivors and long-survivors.

Keywords: survival prediction, brain tumor segmentation, 3D CNN, multimodal MRI, deep learning

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Adriano Pinto,
University of Minho, Portugal
Dong-Hoon Lee,
University of Sydney, Australia

***Correspondence:**
Lin Luo
luol@pku.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *26 April 2019*
Accepted: *22 July 2019*
Published: *16 August 2019*

Citation:
Sun L, Zhang S, Chen H and Luo L (2019) Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning. Front. Neurosci. 13:810. doi: 10.3389/fnins.2019.00810

1. INTRODUCTION

A brain tumor is a cancerous or noncancerous mass or growth of abnormal cells in the brain. Originating in the glial cells, gliomas are the most common brain tumor (Ferlay et al., 2010). Depending on the pathological evaluation of the tumor, gliomas can be categorized into glioblastoma (GBM/HGG), and lower grade glioma (LGG). Glioblastoma is one of the most aggressive and fatal human brain tumors (Bleeker et al., 2012). Gliomas contain various heterogeneous histological sub-regions, including peritumoral edema, a necrotic core, an enhancing and a non-enhancing tumor core. Magnetic resonance imaging (MRI) is commonly used in radiology to portray the phenotype and intrinsic heterogeneity of gliomas, since multimodal MRI scans, such as T1-weighted, contrast enhanced T1-weighted (T1Gd), T2-weighted, and Fluid Attenuation Inversion Recovery (FLAIR) images, provide complementary profiles for different sub-regions of gliomas. For example, the enhancing tumor sub-region is described by areas that show hyper-intensity in a T1Gd scan when compared to a T1 scan.

Accurate and robust predictions of overall survival, using automated algorithms, for patients diagnosed with gliomas can provide valuable guidance for diagnosis, treatment planning, and outcome prediction (Liu et al., 2018). However, it is difficult to select reliable and potent prognostic features. Medical imaging (e.g., MRI, CT) can provide radiographic phenotype of tumor, and it has been exploited to extract and analyze quantitative imaging features (Gillies et al., 2016). Clinical

data, including patient age and resection status, can also provide important information about patients' outcome.

Segmentation of gliomas in pre-operative MRI scans, conventionally done by expert board-certified neuroradiologists, can provide quantitative morphological characterization and measurement of glioma sub-regions. It is also a pre-requisite for survival prediction since most potent features are derived from the tumor region. This quantitative analysis has great potential for diagnosis and research, as it can be used for grade assessment of gliomas and planning of treatment strategies. But this task is challenging due to the high variance in appearance and shape, ambiguous boundaries and imaging artifacts, while automatic segmentation has the advantage of fast speed, consistency in accuracy and immunity to fatigue (Sharma and Aggarwal, 2010). Until now, the automatic segmentation of brain tumors in multimodal MRI scans is still one of the most difficult tasks in medical image analysis. In recent years, deep convolutional neural networks (CNNs) have achieved great success in the field of computer vision. Inspired by the biological structure of visual cortex (Fukushima, 1980), CNNs are artificial neural networks with multiple hidden convolutional layers between the input and output layers. They have non-linear properties and are capable of extracting higher level representative features (Gu et al., 2018). Deep learning methods with CNN have shown excellent results on a wide variety of other medical imaging tasks, including diabetic retinopathy detection (Gulshan et al., 2016), skin cancer classification (Esteva et al., 2017), and brain tumor segmentation (Çiçek et al., 2016; Isensee et al., 2017; Wang et al., 2017; Sun et al., 2018).

In this paper, we present a novel deep learning-based framework for segmentation of a brain tumor and its subregions from multimodal MRI scans, and survival prediction based on radiomic features extracted from segmented tumor sub-regions as well as clinical features. The proposed framework for brain tumor segmentation and survival prediction using multimodal MRI scans consists of the following steps, as illustrated in **Figure 1**. First, tumor subregions are segmented using an ensemble model comprising three different convolutional neural network architectures for robust performance through voting (majority rule). Then radiomic features are extracted from tumor sub-regions and total tumor volume. Next, decision tree regression model with gradient boosting is used to fit the training data and rank the importance of features based on variance reduction. Cross validation is used to select the optimal number of top-ranking features to use. Finally, a random forest regression model is used to fit the training data and predict the overall survival of patients.

2. MATERIALS AND METHODS

2.1. Dataset

We utilized the BraTS 2018 dataset (Menze et al., 2015; Bakas et al., 2017a,b,c, 2018) to evaluate the performance of our methods. The training set contained images from 285 patients, including 210 HGG and 75 LGG. The validation set contained MRI scans from 66 patients with brain tumors of an unknown grade. It was a predefined set constructed by BraTS challenge organizers. The test set contained images from 191 patients with a brain tumor, in which 77 patients had a resection state of Gross Total Resection (GTR) and were evaluated for survival prediction. Each patient was scanned with four sequences: T1, T1Gd, T2, and FLAIR. All the images were skull-striped and re-sampled to an isotropic $1mm^3$ resolution, and the four sequences of the same patient had been co-registered. The ground truth of segmentation mask was obtained by manual segmentation results given by experts. The evaluation of the model performance on the validation and testing set is performed on CBICA's Image Processing Portal ipp.cbica.upenn.edu. Segmentation annotations comprise of the following tumor subtypes: Necrotic/non-enhancing tumor (NCR), peritumoral edema (ED), and Gd-enhancing tumor (ET). Resection status and patient age are also provided. The overall survival (OS) data, defined in days, is also included in the training set. The distribution of patients' age is shown in **Figure 2**.

2.2. Data Preprocessing

Since the intensity value of MRI is dependent on the imaging protocol and scanner used, we applied intensity normalization to reduce the bias in imaging. More specifically, the intensity value of each MRI is subtracted by the mean and divided by the standard deviation of the brain region. In order to reduce overfitting, we applied random flipping and random gaussian noise to augment the training set.

2.3. Network Architecture

In order to perform accurate and robust brain tumor segmentation, we use an ensemble model comprising of three different convolutional neural network architectures. A variety of models have been proposed for tumor segmentation. Generally, they differ in model depth, filter number, connection way and others. Different model architectures can lead to different model performance and behavior. By training different kinds of models separately and by merging the results, the model variance can be decreased, and the overall performance can be improved (Polikar, 2006; Kamnitsas et al., 2017). We used three different CNN models and fused the result by voting (majority rule). The detailed description of each model will be discussed in the following sections.

2.3.1. CA-CNN

The first network we employed was Cascaded Anisotropic Convolutional Neural Network (CA-CNN) proposed by Wang et al. (2017). The cascade is used to convert multi-class segmentation problem into a sequence of three hierarchical binary segmentation problems. The network is illustrated in **Figure 3**.

This architecture also employs anisotropic and dilated convolution filters, which are combined with multi-view fusions to reduce false positives. It also employs residual connections (He et al., 2016), batch normalization (Ioffe and Szegedy, 2015) and multi-scale prediction to boost the performance of segmentation. For implementation, we trained the CA-CNN model using Adam optimizer (Kingma and Ba, 2014) and set Dice coefficient (Milletari et al., 2016) as the loss function. We set the initial

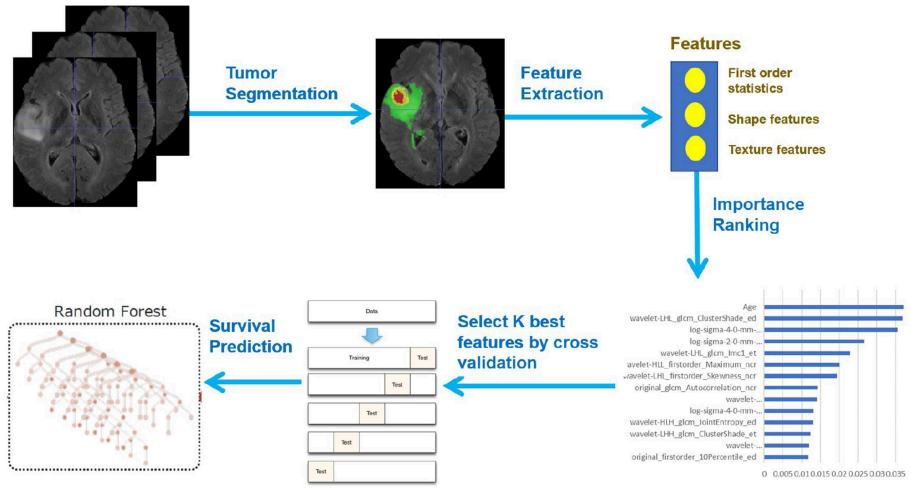


FIGURE 1 | Framework overview.

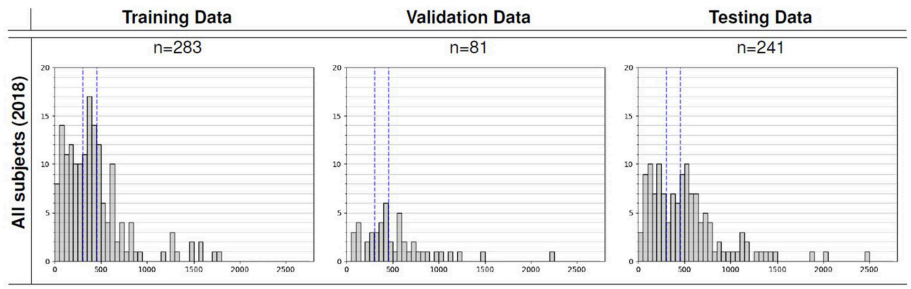


FIGURE 2 | Overall survival distribution of patients across the training, validation, and testing sets.

learning rate to 1×10^{-3}, weight decay 1×10^{-7}, batch size 5, and maximal iteration $30k$.

2.3.2. DFKZ Net

The second network we employed was DFKZ Net, which was proposed by Isensee et al. (2017) from the German Cancer Research Center (DFKZ). Inspired by U-Net, DFKZ Net employs a context encoding pathway that extracts increasingly abstract representations of the input, and a decoding pathway used to recombine these representations with shallower features to precisely segment the structure of interest. The context encoding pathway consists of three content modules, each has two $3 \times 3 \times 3$ convolutional layers and a dropout layer with residual connection. The decoding pathway consists of three localization modules, each containing $3 \times 3 \times 3$ convolutional layers followed by a $1 \times 1 \times 1$ convolutional layer. For the decoding pathway, the output of layers of different depths are integrated by elementwise summation, thus the supervision can be injected deep in the network. The network is illustrated in **Figure 4**.

For implementation, we trained the network using the Adam optimizer. To address the problem of class imbalance, we utilized the multi-class Dice loss function (Isensee et al., 2017):

$$L = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_i u_{i(k)} v_{i(k)}}{\sum_i u_{i(k)} + \sum_i v_{i(k)}} \quad (1)$$

where u denotes output possibility, v denotes one-hot encoding of ground truth, k denotes the class, K denotes the total number of classes and $i(k)$ denotes the number of voxels for class k in patch. We set initial learning rate 5×10^{-4} and used instance normalization (Ulyanov et al., 2016a). We trained the model for 90 epochs.

2.3.3. 3D U-Net

U-Net (Ronneberger et al., 2015; Çiçek et al., 2016) is a classical network for biomedical image segmentation. It consists of a contracting path to capture context and a symmetric expanding path that enables precise localization with extension. Each pathway has three convolutional layers with dropout and pooling. The contracting pathway and expanding pathway are linked by skip-connections. Each layer contains $3 \times 3 \times 3$ convolutional kernels. The first convolutional layer has 32 filters, while deeper layers contains twice filters than previous shallower layer.

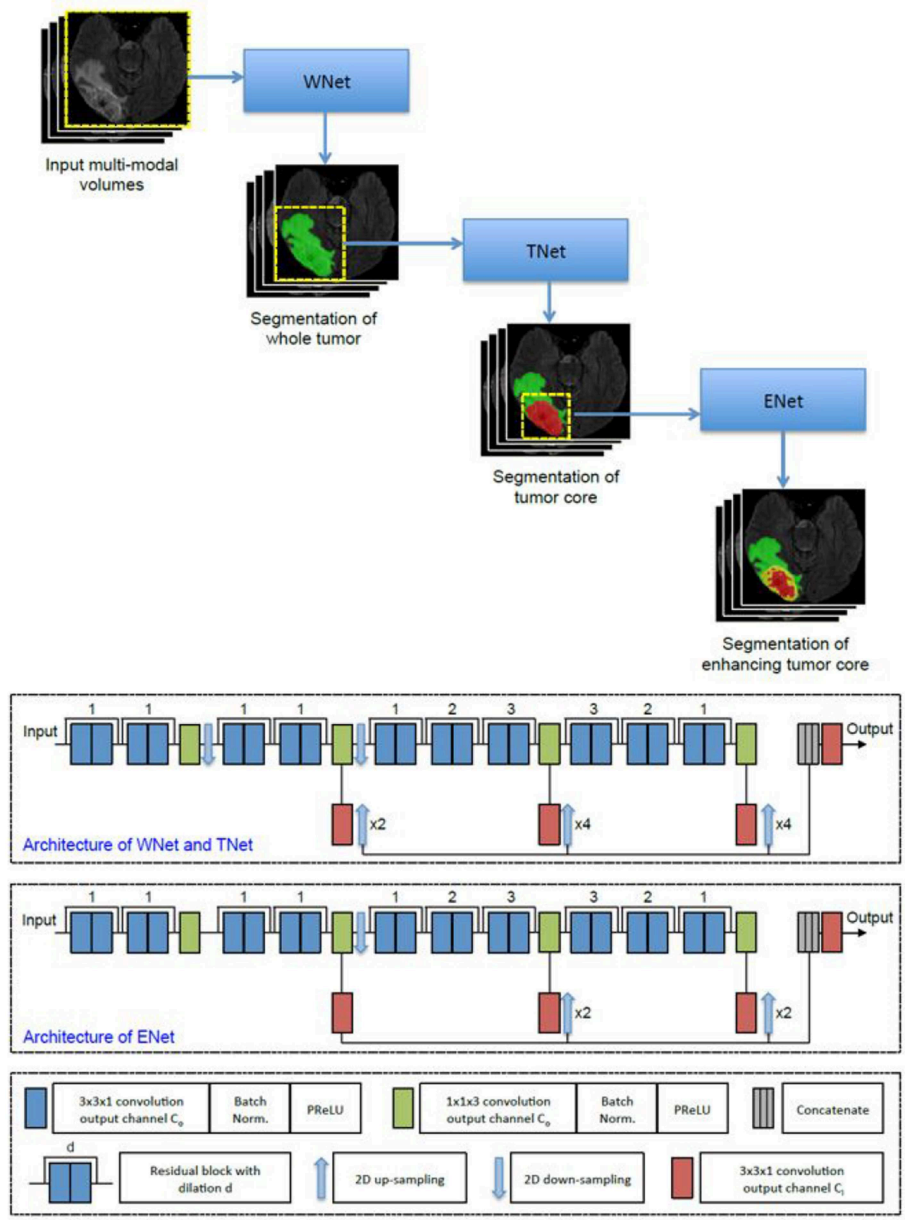


FIGURE 3 | Cascaded framework and architecture of CA-CNN.

For implementation, we used Adam optimizer (Kingma and Ba, 2015), and instance normalization (Ulyanov et al., 2016b). In addition, we utilized cross entropy as the loss function. The initial learning rate was 0.001, and the model is trained for 4 epochs.

2.3.4. Ensemble of Models

In order to enhance segmentation performance and to reduce model variance, we used the voting strategy (majority rule) to build an ensemble model without using a weighted scheme. During the training process, different models were trained independently. The selection of the number of iterations in the training process was based on the model's performance in the validation set. In the testing stage, each model independently predicts the class for each voxel, the final class is determined by the majority rule.

2.4. Feature Extraction

Quantitative phenotypic features from MRI scans can reveal the characteristics of brain tumors. Based on the segmentation result, we extract radiomics features from edema, non-enhancing solid core and necrotic/cystic core and the whole tumor region respectively using *Pyradiomics* toolbox (Van Griethuysen et al., 2017). Illustration of feature extraction is shown in **Figure 5**.

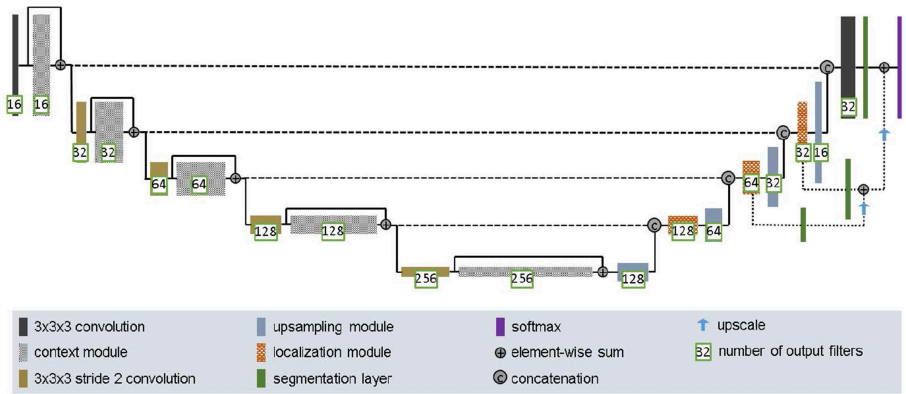


FIGURE 4 | Architecture of DFKZ Net.

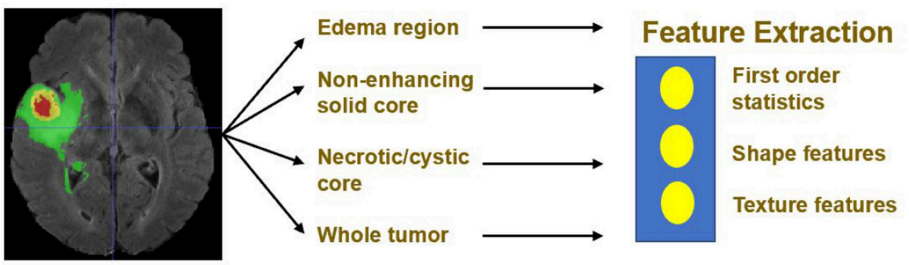


FIGURE 5 | Illustration of feature extraction.

The modality used for feature extraction is dependent on the intrinsic properties of the tumor subregion. For example, edema features are extracted from FLAIR modality, since it is typically depicted by hyper-intense signal in FLAIR. Non-enhancing solid core features are extracted from T1Gd modality, since the appearance of the necrotic (NCR) and the non-enhancing (NET) tumor core is typically hypo-intense in T1Gd when compared to T1. Necrotic/cystic core tumor features are extracted from T1Gd modality, since it is described by areas that show hyper-intensity in T1Gd when compared to T1.

The features we extracted can be grouped into three categories. The first category is the first order statistics, which includes maximum intensity, minimum intensity, mean, median, 10th percentile, 90th percentile, standard deviation, variance of intensity value, energy, entropy, and others. These features characterize the gray level intensity of the tumor region.

The second category is shape features, which include volume, surface area, surface area to volume ratio, maximum 3D diameter, maximum 2D diameter for axial, coronal and sagittal plane respectively, major axis length, minor axis length and least axis length, sphericity, elongation, and other features. These features characterize the shape of the tumor region.

The third category is texture features, which include 22 gray level co-occurrence matrix (GLCM) features, 16 gray level run length matrix (GLRLM) features, 16 Gray level size zone matrix (GLSZM) features, five neighboring gray tone difference matrix (NGTDM) features and 14 gray level dependence matrix (GLDM) Features. These features characterize the texture of the tumor region.

Not only do we extract features from original images, but we also extract features from Laplacian of Gaussian (LoG) filtered images and images generated by wavelet decomposition. Because LoG filtering can enhance the edge of images, possibly enhance the boundary of the tumor, and wavelet decomposition can separate images into multiple levels of detail components (finer or coarser). More specifically, from each region, 1131 features are extracted, including 99 features extracted from the original image, and 344 features extracted from Laplacian of Gaussian filtered images, since we used four filters with sigma values 2.0, 3.0, 4.0, 5.0, respectively, and 688 features extracted from eight wavelet decomposed images (all possible combinations of applying either a High or a Low pass filter in each of the three dimensions). In total, for each patient, we extracted $1131 \times 4 = 4524$ radiomic features, these features are combined with clinical data (age and resection state) for survival prediction. The values of these features except for resection state are normalized by subtracting the mean and scaling it to unit variance.

2.5. Feature Selection

A portion of the features we extracted were redundant or irrelevant to survival prediction. In order to enhance performance and reduce overfitting, we applied feature selection to select a subset of features that have the most predictive power.

Feature selection is divided into two steps: importance ranking and cross validation. We ranked the importance of features by fitting a decision tree regressor with gradient boosting using training data, then the importance of features can be determined by how effectively the feature can reduce intra-node standard deviation in leaf nodes. The second step is to select the optimal

TABLE 1 | Selected most predicative features (WT, edema; TC, tumor core; ET, enhancing tumor; FULL, full tumor volume comprised of edema, tumor core, and enhancing tumor; N/A, not applicable).

Extracted from	Name	Subregion	Score
clinical	age	N/A	0.037375134
wavelet-LHL	glcm_ClusterShade	WT	0.036912293
log-sigma-4.0mm-3D	glcm_Correlation	TC	0.035558309
log-sigma-2.0mm-3D	gldm_LargeDependenceHighGrayLevelEmphasis	TC	0.026591038
wavelet-LHL	glcm_Informational Measure of Correlation	ET	0.022911978
wavelet-HLL	firstorder_Maximum	ET	0.020121927
wavelet-LHL	firstorder_Skewness	ET	0.019402119
original image	glcm_Autocorrelation	ET	0.014204463
wavelet-HHH	gldm_LargeDependenceLowGrayLevelEmphasis	FULL	0.014085406
log-sigma-4.0mm-3D	firstorder_Mwtian	WT	0.013031814
wavelet-HLH	glcm_JointEntropy	WT	0.013023534
wavelet-LHH	glcm_ClusterShade	TC	0.012335471
wavelet-HLL	glszm_LargeAreaHighGrayLevelEmphasis	FULL	0.011980896
original image	firstorder_10Percentile	WT	0.011803132

TABLE 2 | Evaluation result of ensemble model and individual models.

Stage	Metric	Enhancing tumor	Whole tumor	Tumor core
CA-CNN	Mean Dice	0.77682	0.90282	**0.85392**
	Mean Hausdorff95(mm)	3.3303	**5.41478**	6.56793
	Sensitivity	0.81258	0.93045	0.85305
	Specificity	0.99807	0.99336	0.99786
DFKZ Net	Mean Dice	0.76759	0.89306	0.82459
	Mean Hausdorff95(mm)	5.90781	5.60224	6.91403
	Sensitivity	0.80419	0.89128	0.81196
	Specificity	0.99833	0.99588	0.99849
3D U-Net	Mean Dice	0.78088	0.88762	0.82567
	Mean Hausdorff95(mm)	7.73567	12.63285	13.33634
	Sensitivity	0.84281	0.90188	0.81913
	Specificity	0.99743	0.99416	0.9981
Ensemble model	Mean Dice	**0.80522**	**0.90944**	0.84943
	Mean Hausdorff95(mm)	**2.77719**	6.32753	**6.37318**
	Sensitivity	0.83064	0.90688	0.83156
	Specificity	0.99815	0.99549	0.99863

The bold values indicate the best performance.

TABLE 3 | Evaluation result of ensemble model for segmentation.

Stage	Metric	Enhancing tumor	Whole tumor	Tumor core
Validation	Mean Dice	0.8052	0.9044	0.8494
	Mean Hausdorff95(mm)	2.7772	6.3275	6.3732
Testing	Mean Dice	0.7171	0.8762	0.7977
	Mean Hausdorff95(mm)	4.9782	7.2009	6.4735

number of best features for prediction by cross validation. In the end, we selected 14 features and their importance are listed in **Table 1**. The detailed feature definition can be found at (https://pyradiomics.readthedocs.io/en/latest/features.html), last accessed on 30 June 2018.

Unsurprisingly, age had the most predictive power among all of the features. The rest of the features selected came from both original images and derived images. We also found that most features selected came from images generated by wavelet decomposition.

2.6. Survival Prediction

Based on the 14 features selected, we trained a random forest regression model (Ho, 1995) for final survival prediction. The random forest regressor is a meta regressor of 100 base decision tree regressors. Each base regressor is trained on a bootstrapped sub-dataset into order to introduce randomness and diversity. Finally, the prediction from base regressors are averaged to improve prediction accuracy, robustness and suppress overfitting. Mean squared error is used as loss function when constructing individual regression model.

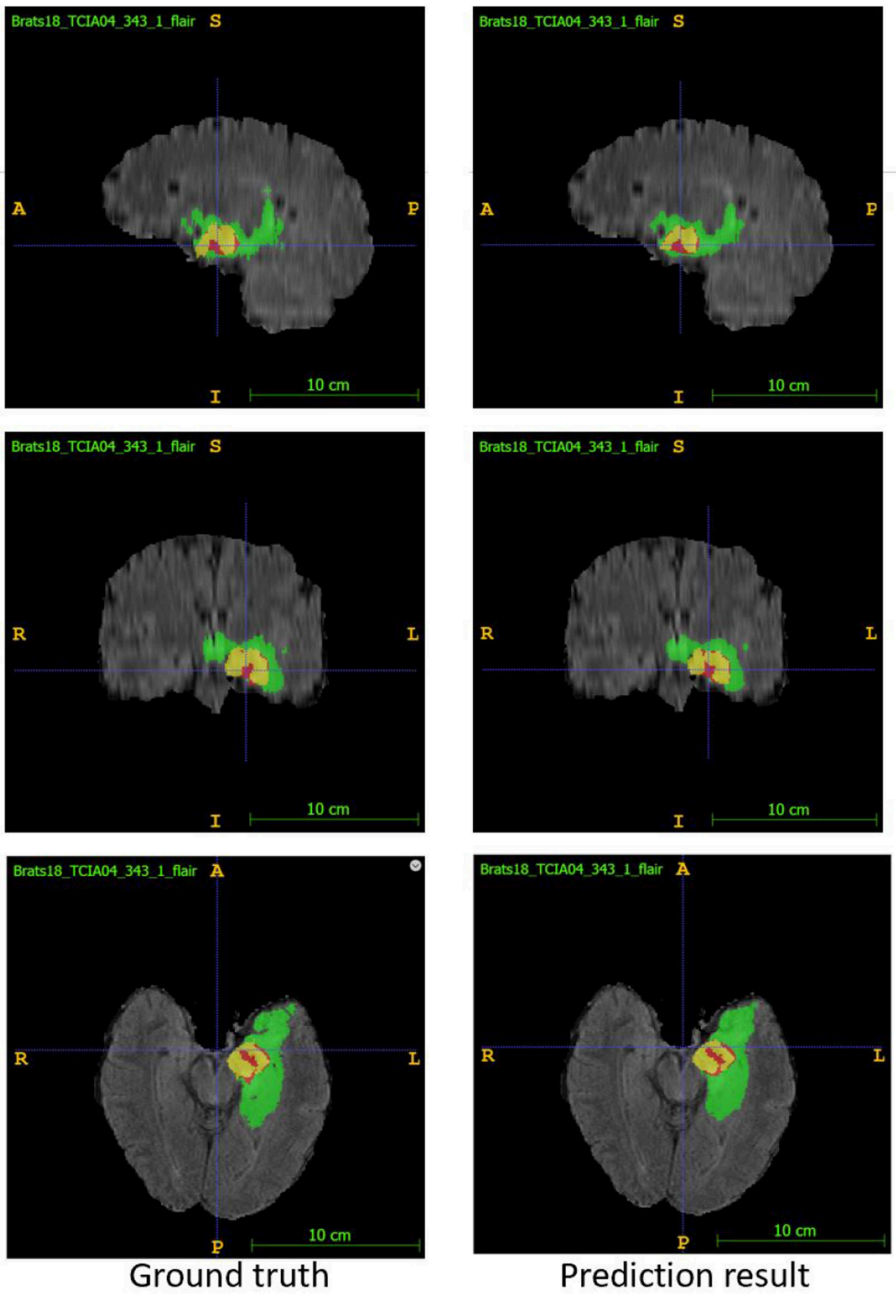


FIGURE 6 | Examples of segmentation result compared with ground truth. Image ID: TCIA04_343_1, Green:edema, Yellow:non-enhancing solid core, Red:enhancing core.

TABLE 4 | Evaluation result of survival prediction.

Stage	Classification accuracy	Median error
Validation	46.4%	217.92
Test	61.0%	181.37

3. RESULTS

3.1. Result of Tumor Segmentation

We trained the model using the 2018 MICCAI BraTS training set using the methods described above. We then applied the trained model for prediction on the validation and test set. We compared the segmentation result of the ensemble model with the individual model on the validation set. The evaluation result of our approach is shown in **Table 2**. For other teams' performance, please see the BraTS summarizing paper (Bakas et al., 2018). The result demonstrates that the ensemble model performs better than individual models in enhancing tumor and whole tumor, while CA-CNN performs marginally better on the tumor core.

The predicted segmentation labels are uploaded to the CBICA's Image Processing Portal (IPP) for evaluation. BraTS Challenge uses two schemes for evaluation: Dice score and the Hausdorff distance (95th percentile). Dice score is a widely used overlap measure for pairwise comparison of segmentation mask S and G. It can be expressed in terms of set operations:

$$Dice = \frac{2|S \cap G|}{|S| + |G|} \tag{2}$$

Hausdorff distance is the maximum distance of a set to the nearest point in the other set, defined as:

$$d_{\mathrm{H}}(X, Y) = \max\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \} \tag{3}$$

where *sup* represents the supremum and *inf* the infimum. In order to have more robust results and to avoid issues with noisy segmentation, the evaluation scheme uses the 95th percentile.

In the test phase, our result ranked 5th out of 60+ teams. The evaluation result of the segmentation on the validation and test set are listed in **Table 3**. Examples of the segmentation result compared with ground truth are shown in **Figure 6**.

3.2. Result of Survival Prediction

Based on the segmentation result of brain tumor subregions, we extracted features from brain tumor sub-regions segmented from MRI scans and trained the survival prediction model as described above. We then used the model to predict patient's overall survival on the validation and test set. The predicted overall survival was uploaded to the IPP for evaluation. We used two schemes for evaluation: classification of subjects as long-survivors (> 15 months), short-survivors (< 10 months), and mid-survivors (between 10 and 15 months) and median error (in days). In the test phase, we ranked second out of 60+ teams. The evaluation results of our method are listed in **Table 4**. For other teams' performance, please see the BraTS summarizing paper (Bakas et al., 2018).

4. DISCUSSION

In this paper, we present an automatic framework for the prediction of survival in glioma using multimodal MRI scans and clinical features. First, a deep convolutional neural network is used to segment a tumor region from MRI scans, then radiomics features are extracted and combined with clinical features to predict overall survival. For tumor segmentation, we used ensembles of three different 3D CNN architectures for robust performance through voting (majority rule). The evaluation results show that the ensemble model performs better than individual models, which indicates that the ensemble approach can effectively reduce model bias and boost performance. Although the Dice score for segmentation is promising, we noticed that the specificity of the model is much higher than the sensitivity, indicating an under-segmentation of the model. For survival prediction, we extracted shape features, first order statistics, and texture features from segmented tumor sub-region, then used a decision tree and cross validation to select features. Finally, a random forest model was trained to predict the overall survival of patients. The accuracy for three-class classification is 61.0%, which still leaves room for improvement. Part of the reason is that we only had a very limited number of samples (285 patients) to train the regression model. In addition, imaging and limited clinical features may only explain patients' survival outcome partially, too. In the future, we will explore different network architectures and training strategies to further improve our result. We will also design new features and optimize our feature selection methods for survival prediction.

DATA AVAILABILITY

The datasets analyzed for this study can be found in the BraTS 2018 dataset https://www.med.upenn.edu/sbia/brats2018/data.html.

AUTHOR CONTRIBUTIONS

LS and SZ performed the analysis and prepared the manuscript. HC helped with the analysis. LL conceived the project, supervised and funded the study, and prepared the manuscript.

FUNDING

Financial support from the Shenzhen Science and Technology Innovation (SZSTI) Commission (JCYJ20180507181527806 and JCYJ20170817105131701) is gratefully acknowledged.

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. *Cancer Imaging Arch.* 286. doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *Cancer Imaging Arch.* 286. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., et Int, and Menze, B. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* arXiv:1811.02629.

Bleeker, F. E., Molenaar, R. J., and Leenstra, S. (2012). Recent advances in the molecular understanding of glioblastoma. *J. Neuro Oncol.* 108, 11–27. doi: 10.1007/s11060-011-0793-0

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, eds S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells (Cham: Springer International Publishing), 424–432.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115–118. doi: 10.1038/nature21056

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C., and Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: Globocan 2008. *Int. J. Cancer* 127, 2893–2917. doi: 10.1002/ijc.25516

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* 36, 193–202.

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: images are more than pictures, they are data. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. *Patt. Recogn.* 77, 354–377. doi: 10.1016/j.patcog.2017.10.013

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410. doi: 10.1001/jama.2016.17216

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Beijing), 770–778.

Ho, T. K. (1995). "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95 (Washington, DC: IEEE Computer Society), 278.

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv* arXiv:1502.03167.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2017). "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge," in *International MICCAI Brainlesion Workshop* (Springer), 287–297.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2017). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI Brainlesion Workshop* (London: Springer), 450–462.

Kingma, D. P. and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv* arXiv:1412.6980.

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations* (Amsterdam).

Liu, L., Zhang, H., Wu, J., Yu, Z., Chen, X., Rekik, I., et al. (2018). Overall survival time prediction for high-grade glioma patients based on large-scale brain functional networks. *Brain Imaging Behav.* 1–19. doi: 10.1007/s11682-018-9949-2

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imag.* 34:1993–2024. doi: 10.1109/TMI.2014.2377694

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, (Munich: IEEE) 565–571.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circ. Syst. Magaz.* 6, 21–45. doi: 10.1109/MCAS.2006.1688199

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image computing and Computer-Assisted Intervention* (Freiburg: Springer), 234–241.

Sharma, N., and Aggarwal, L. M. (2010). Automated medical image segmentation techniques. *J. Med. Phys. Assoc. Med. Phys. India* 35:3–14. doi: 10.4103/0971-6203.58777

Sun, L., Zhang, S., and Luo, L. (2018). "Tumor segmentation and survival prediction in glioma with deep learning," in *International MICCAI Brainlesion Workshop* (Shenzhen: Springer), 83–93.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016a). Instance normalization: The missing ingredient for fast stylization. *arXiv* arXiv:1607.08022.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016b). Instance normalization: the missing ingredient for fast stylization. arxiv 2016. *arXiv* arXiv:1607.08022.

Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (London: Springer), 178–190.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 27 August 2019
doi: 10.3389/fncom.2019.00058

A Multi-parametric MRI-Based Radiomics Signature and a Practical ML Model for Stratifying Glioblastoma Patients Based on Survival Toward Precision Oncology

*Alexander F. I. Osman**

Department of Medical Physics, Al-Neelain University, Khartoum, Sudan

Purpose: Predicting patients' survival outcomes is recognized of key importance to clinicians in oncology toward determining an ideal course of treatment and patient management. This study applies radiomics analysis on pre-operative multi-parametric MRI of patients with glioblastoma from multiple institutions to identify a signature and a practical machine learning model for stratifying patients into groups based on overall survival.

Methods: This study included 163 patients' data with glioblastoma, collected by BRATS 2018 Challenge from multiple institutions. In this proposed method, a set of 147 radiomics image features were extracted locally from three tumor sub-regions on standardized pre-operative multi-parametric MR images. LASSO regression was applied for identifying an informative subset of chosen features whereas a Cox model used to obtain the coefficients of those selected features. Then, a radiomics signature model of 9 features was constructed on the discovery set and it performance was evaluated for patients stratification into short- (<10 months), medium- (10–15 months), and long-survivors (>15 months) groups. Eight ML classification models, trained and then cross-validated, were tested to assess a range of survival prediction performance as a function of the choice of features.

Results: The proposed mpMRI radiomics signature model had a statistically significant association with survival ($P < 0.001$) in the training set, but was not confirmed ($P = 0.110$) in the validation cohort. Its performance in the validation set had a sensitivity of 0.476 (short-), 0.231 (medium-), and 0.600 (long-survivors), and specificity of 0.667 (short-), 0.732 (medium-), and 0.794 (long-survivors). Among the tested ML classifiers, the ensemble learning model's results showed superior performance in predicting the survival classes, with an overall accuracy of 57.8% and AUC of 0.81 for short-, 0.47 for medium-, and 0.72 for long-survivors using the LASSO selected features combined with clinical factors.

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Gaurav Shukla,
University of Pennsylvania,
United States
Ahmad Chaddad,
McGill University Health Centre, Canada

***Correspondence:**
Alexander F. I. Osman
alexanderfadul@yahoo.com

Received: *30 April 2019*
Accepted: *09 August 2019*
Published: *27 August 2019*

Citation:
Osman AFI (2019) A Multi-parametric MRI-Based Radiomics Signature and a Practical ML Model for Stratifying Glioblastoma Patients Based on Survival Toward Precision Oncology. Front. Comput. Neurosci. 13:58. doi: 10.3389/fncom.2019.00058

Conclusion: A derived GLCM feature, representing intra-tumoral inhomogeneity, was found to have a high association with survival. Clinical factors, when added to the radiomics image features, boosted the performance of the ML classification model in predicting individual glioblastoma patient's survival prognosis, which can improve prognostic quality a further step toward precision oncology.

Keywords: glioblastoma multiforme, MRI, radiomics analysis, patient's survival prediction, machine learning, precision oncology

INTRODUCTION AND RELATED WORKS

Introduction

Glioblastoma multiforme (GBM) is the most aggressive and highly invasive high-grade glioma tumors with poor prognosis (Holland, 2001). The median survival rate of GBM patients is about 2 years or less, and it needs immediate treatment (Ohgaki and Kleihues, 2005; Louis et al., 2007). Surgical resection followed by chemo-radiotherapy is the current standard treatment of the glioblastoma multiforme tumors (Van Meir et al., 2010; Aum et al., 2014). Predicting a patient's survival outcome is recognized as key importance to clinicians in oncology toward determining an ideal course of treatment and patient management. In which, the treating physician (oncologist) may decide if more aggressive or additional treatment has to be considered for treating patients with poor survival prognosis (Zhang et al., 2017).

Multi-parametric magnetic resonance imaging (mpMRI) sequences commonly provide more clinical information to characterize glioblastoma multiforme tumors than other imaging modalities. Here, "multi-parametric" is refereed to multiple image standardization parameters. This imaging information could be quantitatively extracted as features and linking these tumor phenotype features to clinical variables of interest (e.g., survival time, recurrence, adverse events, or late complications). The mentioned concept is referred to as radiomics. The idea of radiomics has recently emerged from the field of oncology. Radiomics has the potential for enabling improved clinical decision-making (Gillies et al., 2016). This approach has advantages of being non-invasive, fast and low in cost. Radiomics has been used in oncology for tumors' diagnosis, treatment planning/execution, treatment response and prognosis, and underlying genomic patterns in various forms of cancer (Liu et al., 2018a). In which, individual patients could be stratified into subtypes based on radiomics biomarkers that hold information about cancer traits that reflect the patient's prognosis. As a result, radiomics could have an effective application in precision oncology by predicting individual patients' treatment outcome.

The definition of precision medicine, according to the National Institute of Health (NIH), is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person" (Subramaniam, 2017). This concept will let clinicians and researchers provide predictions with higher accuracy for which treatment and prevention plans for a particular disease will suit in which groups of people (Subramaniam, 2017). The newly introduced idea of precision medicine is in contrast to the existing practical therapy paradigm of a "one-size-fits-all" attitude, in which disease treatment and prevention plans are developed for the "average" patient, with less consideration for the differences between individuals (Subramaniam, 2017). There are some limitations in fully implementing precision medicine for radiomics e.g., reproducibility and quantitative information, standardization in image acquisition, and structured reporting.

The Related Works

Many studies have been conducted identifying tumor phenotypical radiomics signature or/and developing practical machine learning (ML) models for glioblastoma patients stratification based on survival on pre-operative multi-parametric MRI sequences from single or multiple institutions. Recognizing patients who would/wouldn't benefit from standard treatment as well as identifying patients who need more aggressive treatment at the time of diagnosis is essential toward management of glioblastoma through personalized medicine. In this section, the author included some works of the most relevant ones recently published in this field. Macyszyn et al. (2016) used image analysis and ML models to establish imaging patterns that are predictive of overall survival (OS) and molecular subtype using preoperative mpMRIs sequences of patients with GBM. The developed system achieved an overall accuracy of 80% in stratifying patients into long-, medium-, and short-term survivors in the prospective cohort from a single institution. Prasanna et al. (2017) studied texture features analysis to assess the efficacy of peritumoral brain zone features from pre-operative MRI in predicting GBM patient survival into long- (>18 months) vs. short-term (<7 months). The study findings identified a subset of 10 features proven to be predictive of long- vs. short-term survival as compared to known clinical factors. Ingrisch et al. (2017) investigated whether radiomics analysis with random survival forests can predict overall survival from MRI scans of newly diagnosed glioblastoma patients. Their results demonstrated that low predicted individual mortality proven to be a favorable prognostic factor for OS, it also indicated that the MRI contains prognostic information, which can be accessed by radiomics analysis.

Most recently, Chaddad et al. (2018) proposed multiscale texture features for predicting GBM patients' progression-free survival and overall survival on T1 and T2-FLAIR MRIs using the random forest. The study results showed that the identified seven-feature set, when combined with clinical factors, improved the model performance yielding an AUC value of 85.54% for OS predictions. Kickingereder et al. (2018) investigated the

impact of mpMRI radiomics features for predicting patients' survival in newly diagnosed GBM patients before treatment. The study results revealed that a constructed eight-feature radiomics signature increased the prediction accuracy for OS further than the alternative approaches. Sanghani et al. (2018) studied survival prediction of glioblastoma patients for two-class (short- vs. long-term) and three-class (short-, medium-, and long-term) survival groups using Support Vector Machines (SVMs). The results showed a prediction accuracy of 98.7 and 88.95% for two-class and three-class OS group, respectively. Chen et al. (2019) studied developing a post-T1-weighted MRI-based prognostic radiomics classification system in GBM patients to assess if it could allow stratifying patients into a low- or high-risk group. Their results showed that the developed system classified patients' survival with improved performance with AUC of 0.851 for 12-month survival, compared to conventional risk models.

The majority of those studies have performed on single-institution data, and also survival grouping was designed for two-class rather than three-class approach. Besides, implementing a particular feature selection method and testing various machine learning classification models allow greater flexibility for exploring distinct methods. The purpose of this work is to quantitatively study the radiomics features from pre-operative multi-parametric MRI of the *de novo* glioblastoma tumor on multi-institutional datasets. Then, to apply radiomics analysis on mpMRI to identify a signature and a practical machine learning model to stratify patients into short-, medium, and long-survivors groups. For machine learning, different models were tested to assess a range of performance as a function of the choice of features.

MATERIALS AND METHODS

Patients Data Sets

The study involved a cohort of 163 patients diagnosed with primary *de novo* GBM and pathologically confirmed. The patients' imaging data sets and clinical information data were collected from multiple ($n = 5$) institutions and provided as "training data set" for Multimodal Brain Tumor Segmentation (BRATS) 2018 Challenge (Menze et al., 2015; Bakas et al., 2017a,b,c). For each patient, the imaging data set consisted of four sequences of pre-operative multi-parametric MRIs along with the patient's clinical information. The imaging data sets were acquired during regular clinical routine using various scanners, and different scanning protocols. An individual patient's imaging data set included T1-weighted (T1), T1-weighted with post-contrast/gadolinium (T1-Gd), T2-weighted (T2), and T2-weighted fluid-attenuated inversion recovery (T2-FLAIR) MRI sequences. Besides, "ground truth" segmentation masks of three tumor sub-structures provided as follow: the complete tumor extent also referred to as the "whole tumor" (WT), tumor core (TC), and the active tumor (AT) and the non-enhancing/necrotic tumor region (**Figure 1**). The clinical data were composed of the patient's age, patient's overall survival, and tumor's resection status information. The demographic and clinical characteristics data of the glioblastoma patients in the discovery, validation, and in the combined cohorts, were presented in **Table 1**.

The patient data sets were categorized into discovery/training and validation cohorts. In which, the survival data were sorted in order hence after every two consecutive values the third one was chosen for validation and added to the validation data set while the remained ones were considered as the discovery data set. This distribution of overall survival data across the discovery and validation data sets ensure a balanced appearance of the whole OS values range (from short, through a medium, to long-survivors) in both cohorts. The patients' survival data were categorized into long- (>15 months), medium- (between 10 and 15 months), and short-term survivors (<10 months) groups. The reason behind choosing these thresholds can be found with a detailed explanation by referring to this BRATS paper (Bakas et al., 2019).

Annotation of Tumor Structures

The extracted radiomics features may suffer from the robustness due to variations in the delineated tumor structures. Consequently, a decision was made to use the provided "ground truth" segmentation masks which were manually generated by experts, rather than using the author's developed automated segmentation system (Osman, 2018) which was still under further improving. The tumor sub-structures delineation was performed by experts (one to four raters) using the multi-parametric MR images following a specific given annotation protocol. The experts' annotations were further revised by an experienced board-certified neuroradiologist to minimize inter- and intra-raters variations (Menze et al., 2015; Bakas et al., 2019). Three tumor sub-structures were delineated on the imaging data namely; the complete tumor extent also referred to as the "whole tumor," the tumor core, and the active tumor and the non-enhancing/necrotic tumor region structures (illustrated in **Figure 1**). The protocol used for annotating the tumor structures was described in detail in those two BRATS papers (Menze et al., 2015; Bakas et al., 2019).

Image Preprocessing

The multi-parametric MR images were provided with initial preprocessing. The four mpMRI sequences of each patient were co-registered using T1-Gd image sequence as a reference. The images were also smoothed, interpolated to the same resolution of 1 mm^3, and skull-stripped. Each imaging sequence was had 240 × 240 pixels and 155 slices acquisition matrices and converted into grayscale. Further preprocessing were performed to standardize the image intensity before performing features extraction. The most commonly used MRI normalization scheme of $\mu \pm 3\sigma$ with 256 intensity bins (Collewet et al., 2004) was applied. MRI intensity rescaling (**Figure 2**) on the global brain image volume was employed to convert MRI signal intensity values into a standardized intensity range, thus avoiding bias due to heterogeneity. Image intensities were standardized between $\mu \pm 3\sigma$ where μ was the mean value of the gray levels inside the region of interest (brain) and σ the standard deviation. The gray level values outside the $[\mu - 3\sigma,\ \mu + 3\sigma]$ range were truncated to

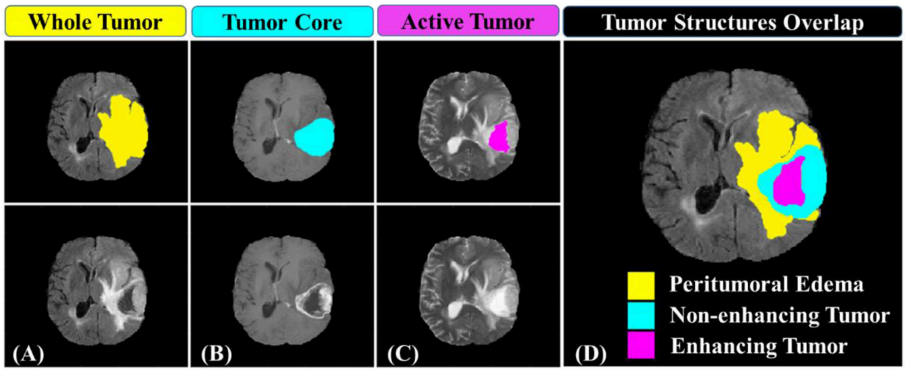


FIGURE 1 | Glioblastoma multiforme sub-regions segmentation masks generated by experts annotated in the different MRI sequences. **(A)** the whole tumor (yellow) visible in T2-FLAIR, **(B)** the tumor core (light blue) visible in T2, and **(C)** the active tumor structures (purple) visible in T1-Gd. Combination of three segmentation labels overlaid on T2-FLAIR MRI producing **(D)** the final labels of the tumor sub-structures: peritumoral edema [ED] (yellow), non-enhancing solid core tumor [NET] (light blue), necrosis [NCR], and enhancing tumor core (purple).

TABLE 1 | Demographic and clinical characteristics data of GBM patients in discovery, validation, and combined sets.

Characteristic	Discovery	Validation	Combined
Patients demographic			
No. of patients			
Patient distribution	109 (67%)	54 (33%)	163
- CBICA UPenn	–	–	85 (52%)
- TCIA	–	–	76 (47%)
- MGH, HU, DU, and BU	–	–	2 (1%)
Imaging data			
- Data set of T1, T1-Gd, T2, and T2-FLAIR MRI sequences with tumor sub-structures "ground truth" segmentation labels	–	–	163
Clinical information			
Age (years) ($P = 0.368$)†			
- Range	18.97–84.84	33.88–85.76	18.97–85.76
- Mean	59.73	61.55	60.33
- Median	60.94	62.36	61.17
- 1 Standard deviation	12.23	11.81	12.03
Overall survival (days) ($P = 0.934$)†			
- Range	5–1767	22–1731	5–1767
- Mean	421.37	426.18	422.96
- Median	362.00	364.50	362.00
- 1 Standard deviation	350.00	352.31	349.67
- Short-term survivors [<10 months]	44	21	65 (40%)
- Medium-term survivors [10–15 months]	28	14	42 (26%)
- Long-term survivors [>15 months]	37	19	56 (34%)
Resection status ($P = 0.474$)†			
- Gross total resection	36	23	59 (36%)
- Subtotal resection	19	5	24 (15%)
- Missing information	54	26	80 (49%)

CBICA UPenn, Center for Biomedical Image Computing and Analytics at the University of Pennsylvania; TCIA, The Cancer Imaging Archive; BU, Bern University; DU, Debrecen University; HU, Heidelberg University; MGH, Massachusetts General Hospital.
† Data in parentheses are P-value.

the upper or lower limit value. The given range was then quantized into 8 bits [0, 255]. This standardization method eliminates the dependency on the shift of the mean value and multiplicative change in the image intensity. In contrast, the relative difference between two gray levels is not maintained (Collewet et al., 2004).

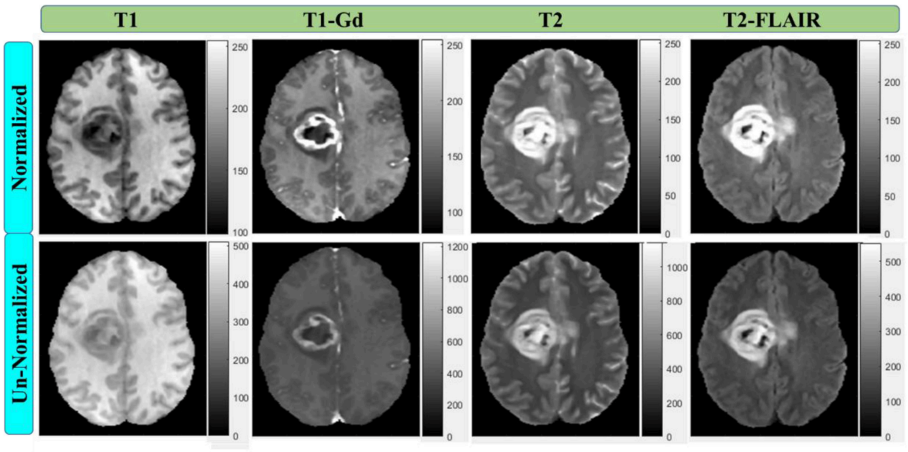


FIGURE 2 | Multi-parametric MRI sequences before and after intensity normalization with 256 scale (8-bit depth). All normalized images have the same scale.

Feature Extraction and Selection

Feature Extraction

For each patient, various features were extracted locally from the "whole tumor", tumor core, and active tumor areas on the T1-Gd, T2, and T2-FLAIR MRI sequences to capture different phenotypic characteristics of the tumor. The features were divided into the following groups:

- Geometry/shape features: which describe the two-dimensional (2D) and 3D shape characteristics of the tumor.
- Intensity features: which describe the first-order statistical distribution of the voxel intensities obtained from a histogram characterizing heterogeneity without giving spatial information within a tumor.
- Gray-Level Co-occurrence Matrix (GLCM) Texture features: which describe the high-order statistical spatial distributions of the voxel intensities characterizing heterogeneity with spatial information within a tumor or region of interest (Haralick et al., 1973; Haralick and Shapiro, 1992).
- Histogram of Oriented Gradients (HOGs) features: which capture local shape information from regions or point locations within an image (Dalal and Triggs, 2005).
- Local Binary Pattern (LBP) features: which encode local texture information that can be used for tasks such as detection and recognition (Ojala et al., 2002).

The normalized volumetric MRI data were used for 2D and 3D features extraction. The 2D features were extracted from a region of interest on a pre-selected image slice. This slice was chosen to correspond to the largest tumor surface area in axial, sagittal, and coronal planes. Then the transverse slice was picked out for extracting the information. Based on the segmentation results [WT on (T2-FLAIR), TC on (T2), or AT on (T1-Gd)], the region enclosing each tumor sub-structure was cropped down on the image. The obtained image was used to extract feature information. A total of 147 multi-parametric MRI radiomics features were extracted/derived for each patient from the segmented tumor sub-structures on the three mpMRI sequences for their capability to characterize the glioblastoma tumor phenotypes. For every sub-region, a set of 48 radiomics features was obtained, resulting in a total of 144 features for the three regions plus 3 additional ones calculated as a joint of the three regions. The features included 14 geometry/shape (plus 3 mixed) features, 14 statistical intensity features, 14 texture (GLCM) features, and 6 local features representing 3 HOG features and 3 LBP features (listed in **Table 2**). All features were derived using MATLAB 2016b Toolbox (Mathworks, Natick, MA, USA) with Image Processing and Computer Vision Tools.

Feature Selection

Following the feature extraction, a feature selection method is required to lessen the number of features to consider only the significant ones. Feature selection refers to reduction of the number of parameters to avoid overfitting dilemma while improving the generalizability and interpretability of the training-based model. Accordingly, a two-step method was applied to choose the most important features and throughout the less associated ones. *Initially*, the median absolute deviations (MAD) was calculated for the 147 extracted features. None of the features with MAD equal to zero, which considered as non-informative, was observed in the total set to be discarded. After this step, the number of features remained the same. *Then*, least absolute shrinkage and selection operator (LASSO) generalized linear regression (Tibshirani, 1996) was employed for finding a subset of the most relevant features from the initial set. Basically, LASSO executes a penalty on the log partial likelihood (sum of squares) that is equal to the absolute sum of regression coefficients. Cross-validation the deviance is then used to determine the LASSO tuning parameter λ (Hastie et al., 2009). LASSO minimizes the regression coefficients down toward zero while it makes the coefficients exactly zero for irrelevant features (Collewet et al., 2004). The LASSO method has been used extensively in high-dimensional feature selection when the number of variables exceeds the sample size (Heinze et al., 2018) as a case in

TABLE 2 | A summary of radiomics features extracted from the tumor sub-regions (WT, TC, and AT) in multi-parametric MR images (T1-Gd, T2, and T2-FLAIR).

Feature classes	Feature names
Sub-regions ($n = 3$)	Whole tumor (WT), tumor core (TC), and active tumor (AT).
Shape features ($n = 14 + 3$*)	Volume [tumor, brain], volume ratio [tumor/brain, AT/WT*, TC/WT*, AT/TC*], surface area [tumor convex area, tumor filled area, tumor area, brain area], surface area ratio [tumor to brain], eccentricity, orientation, equivalent diameter, solidity, extent, perimeter.
Intensity features ($n = 14$)	Minimum value, maximum value, median value, mean value, range, variance, moment 2nd-order, moment 3rd-order, entropy, kurtosis, root mean square (RMS), skewness, standard deviation, mean absolute deviation (MAD).
Texture features: GLCM ($n = 14$)	Contrast, correlation, energy, homogeneity, (sum) variance, (sum) average, (mean) variance, (mean) autocorrelation, entropy, (sum) entropy2, (difference) entropy2, (sum) variance2, (difference) variance2, range of all GLCM features.
HOG features ($n = 3$)	Sum HOG, median HOG, standard deviation HOG.
LBP features ($n = 3$)	Sum LBP, mean LBP, standard deviation LBP.

All features were extracted from a 2D image except those indicated as volumetric features (3D).
Unless noted with a strike (*), each feature was individually extracted from the "whole tumor" area on T2-FLAIR MRI, tumor core area on T2 MRI, and active tumor area on T1-Gd MRI.
**These features were calculated as combined features from joint of WT, TC, and AT sub-structures.*
Features indicated with (2) were derived from GLCM calculated horizontally (0-degree) and 45-degree rotations.

this study where the number of extracted imaging features ($n = 147$) is higher than the number of patients ($n = 109$) in the discovery set. When the LASSO regression model was applied here, nine features with non-zero coefficients retained from all features' set. To search for an optimal λ., cross-validation with 10-fold was applied, where the final λ value yielded minimum error in cross-validation (**Figure 4**). The selected subset was considered as the final one of the chosen features which will be used to construct the multi-parametric MRI radiomics signature model on the discovery data set ($n = 109$).

Constructing and Validating a Radiomics Signature

Using the LASSO regression selected imaging features, a multivariate LASSO Cox regression (Cox and Oakes, 1984) was then applied to obtain the coefficients of those chosen features rather than using the LASSO's coefficients. The reason for using LASSO Cox regression, because it enables getting the *p*-value, and interferes with the coefficients (Tibshirani, 1997). Cox regression is a semiparametric method for fitting survival rate estimates to eliminate the effect of confounding features, and to quantify the effect of predictor features. It has been reported that the LASSO Cox regression model is reliable for prediction of patients' survival in glioma (Chaddad et al., 2019a). The selected image features with their corresponding coefficients were used to construct a mpMRI radiomics signature model. At first, a radiomics risk score for each patient was determined by linearly combining these selected features weighed by their respective fitting coefficients (β) (Liu et al., 2018a) as follows:

$$Risk\ score = \sum_{i=1}^{n} \beta_i \,.\, feature_i.$$

Then, the risk scores obtained for patients in the discovery set were stratified into low-(long-), medium-(medium-), and high-risk (short-survivors), with fixed cutoff points as thresholds. The steps which the author implemented to find these cutoffs were as following: first, the radiomics risk score was calculated for all patients in the discovery set. Their values ranged between (+)4.118 to (–)1.497 for the short-survivors group (high risk), (+)0.945 to (–)2.619 for the medium-survivors group (medium risk), and (+)1.603 to (–)3.211 for the long-survivors group (low risk). Then, the corresponding median (50 percentile) values for each survivor group were determined to be (+)0.245, (–)0.810, and (–)1.009, respectively. Finally, since there was an overlap between the three regions, the author calculated the 25 percentile values (approximated as the half median values) of the high-risk (+0.122) and low-risk (−0.505). Accordingly, these values were used as fixed thresholds for stratifying patient into low-risk (Rad-score <– 0.505) for long-survivors (> 15 months) group, medium-risk (Rad-score between −0.505 and 0.122) for medium-survivors (10–15 months) group, and high-risk (Rad-score > 0.122) for short-survivors (<10 months) group.

The mpMRI radiomics signature model was constructed on the discovery data set. Its statistical performance with survival association was assessed in the discovery and validation sets using the *t*-test. True positive rate (sensitivity) and the false positive rate (1—specificity) metrics were used to evaluate the signature model's classification performance in both data sets. The association between the LASSO selected radiomics features and survival in the discovery and validation data sets was illustrated via a heat map, in which the selected radiomics features were rescaled by the z-score transformation.

Training and Validating a ML Classifier

Several machine learning classification algorithms were assessed in this study for patients' stratification based on survival. The classifiers were trained, and the top-ranked ones reported. Eight various models were included here, and they are listed below:

(A) *Support Vector Machine classifiers* (Vapnik, 1982):

1) *Linear SVM*: makes a basic linear separation of classes;
2) *Medium Gaussian SVM*: creates moderate distinctions between classes, with a kernel scale set to the square root of (P) where P is the number of features/predictors;

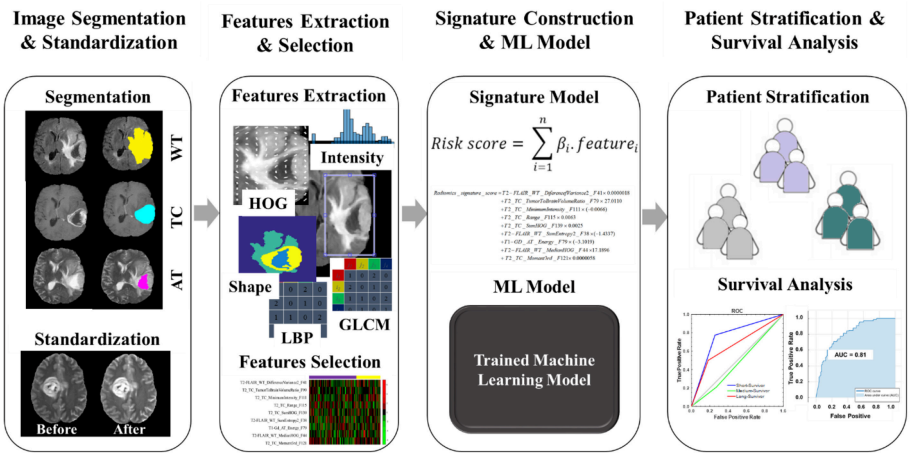


FIGURE 3 | The workflow of radiomics analysis used in this study. The overall procedure of identifying a mpMRI radiomics signature model and a practical ML model for stratifying the GBM patient's prognoses based on overall survival.

3) *Coarse Gaussian SVM*: creates coarse distinctions between classes, with kernel scale set to the square root of (P) × 4,

(B) *K-Nearest Neighbors (KNN) classifiers* (Patrick and Fischer, 1970):

4) *Coarse KNN*: creates rough distinctions between classes with the number of neighbors set to 100;
5) *Cosine KNN*: creates moderate distinctions between classes using a cosine distance metric with the number of neighbors set to 10;
6) *Medium KNN*: creates moderate distinctions between classes with the number of neighbors set to 10,

(C) *Discriminant Analysis* (McLachlan, 2004):

7) *Linear Discriminant*: creates linear boundaries between classes, and

(D) *Ensemble Learning*: (Ho, 1998; McLachlan, 2004):

8) *Subspace Discriminant*: Subspace, with Discriminant Analysis, has medium flexibility and good for many predictors with a few hundred learners. Learning rate set to 0.1 is a popular choice for shrinkage.

All classifiers were trained on the combined data set ($n = 163$). They were trained using various feature combinations: (a) the all radiomics ($n = 147$) features, (b) the LASSO selected ($n = 9$) features, (c) and (d) both features combined with the clinical factors (predictors), respectively. The target response for each model was the patients' OS grouped into three classes representing short- (<10 months), medium- (10–15 months), and long-survivors (>15 months).

A cross-validation scheme with 5-fold (to avoid overfitting) was employed to examine the predictive accuracy of the trained ML classification models and help in determining the best model. The method is commonly recommended for a small data set, as in the case of this study (163 observations). The receiving operating characteristics (ROC) curve was used to check model performance after training each classifier. ROC plot, illustrating the performance of the classifier, displays values of the true-positive and false-positive rates for the model under study. The area under the ROC curve (AUC) was used to measure the performance of individual survival group predicted by a classier, and the accuracy metric to evaluate the overall classifier performance in predicting the three groups. Also, the individual classifier's performance as a function of feature choice was assessed to examine its impact on accuracy.

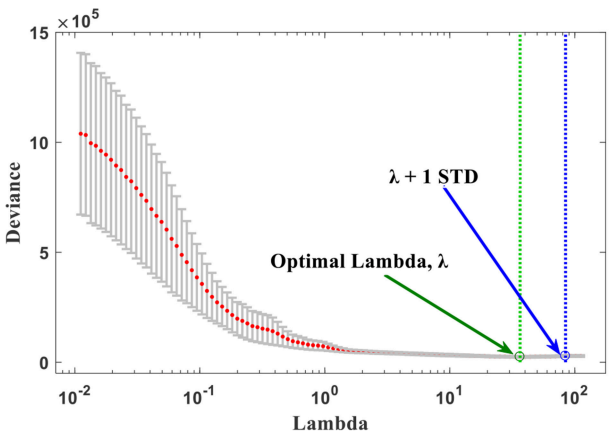


FIGURE 4 | The optimal λ selection by cross-validated deviance of LASSO fit. The partial likelihood deviance plotted vs. λ. The green dotted vertical line was plotted at the optimal λ (36.50) and the blue dotted at λ + 1 STD (84.33) as shown in the plot.

The Proposed Method

The flowchart of the proposed model/method presented in this study for survival prognosis for patients with glioblastoma is demonstrated in **Figure 3**. It composed of four blocks. Block one for image acquisition, segmentation, and preprocessing, block

two for features extraction and selection, block three for signature construction and ML models, and finally block four for patient stratification and survival analysis.

The overall procedure could be summarized as follows: At first, pre-operative multi-parametric MRI (T1, T1-Gd, T2, and T2-FLAIR) sequences are acquired for patients with glioblastoma multiforme (**Figure 1**). Tumor sub-structures ("whole tumor", tumor core, and active tumor) are delineated on the acquired images after registering the images with its corresponding reference one. Then the mpMRI intensities are rescaled with a standardized normalization scheme of $\mu \pm 3\sigma$ with 256 intensity bins (**Figure 2**). Secondly, features extraction and selection take place here. Geometry/shape, intensity, HOG, LPB, and GLCM features (**Table 2**) are derived from the standardized intensity MRIs. Important features with the most relevance to patient survival are selected with LASSO (**Table 3** and **Figure 4**). Thirdly, multivariate LASSO Cox is applied to the selected features to extract the corresponding coefficients. These coefficients are linearly combined to construct a radiomics signature model via risk score. Then, fixed thresholds determined during the signature construction, are used for stratifying patients into a low-risk (Rad-score < -0.505) for long-survivors (>15 months) group, a medium-risk (Rad-score between -0.505 and 0.122) for medium-survivors (10–15 months) group, and a high-risk (Rad-score > 0.122) for short-survivors (<10 months) group. A multivariate ensemble (subspace Discriminant) machine learning model, trained and cross-validated, is used as a more practical model for survival class prediction. And fourthly, using the signature and ML models, glioblastoma individual patients are stratified into short-, medium-, or long-survivors.

Statistical Analysis

All of the statistical data analysis and modeling in this study were performed with MATLAB 2016b software with implemented Statistics and Machine Learning Toolbox (MathWorks, Natick, MA, USA). The differences in patient age, tumor resection status, and OS between the discovery and the validation data sets were evaluated using an independent sample *t*-test (two-sample *t*-test).

RESULTS

Clinical Characteristics

The median and mean of overall survival were 362 days and 421 days for the discovery/training data set. For the validation data set, the values were 364 days and 426 days, respectively. The median and mean of age were 60 years and 61 years, respectively, for the discovery data set, and the values for both, median and mean, were 62 years for the validation data set. There was no indication of significant difference in clinical and follow-up data between the discovery and validation data sets ($P = 0.368$ for age test, $P = 0.474$ for tumor resection status test, and $P = 0.934$ for OS test).

The Radiomics Signature Results

The nine features, selected by the LASSO with non-zero coefficients, formed of 2 from T2-FLAIR, 1 from T1-Gd, and 6 from T2 MRI. These imaging features, plus the clinical factors, are provided in **Table 3**, arranged in order from high to low importance (*P*-value), with their median, *P*-values, and LASSO Cox regression model coefficients. Each feature was named as Modality_Region_FeatureName_FeatureNumber. For instance, T2_TC_SumHOG_F139 indicated that this feature is the sum of HOG extracted from the tumor core region on T2 MRI sequence and was the feature number 139 in the full list. The optimal λ obtained during the cross-validation of features selection in LASSO regression model was 36.50 with $\lambda + 1$ standard deviation (STD) of 84.33 (66.67% confidence level), as shown in **Figure 4**. As a result, this optimized value, obtained through the cross-validation, has selected nine features with non-zero coeffcients. Usually, as the lambda value increases, the number of non-zero components of predictor coefficients decreases.

Features indicated strong association with survival ($P < 0.05$) from most to least, according to their *P*-value as shown in **Table 3** are: GLCM difference variance2 (difference variance calculated at 0 degree and 45 degree rotations) in the WT [T2-FLAIR], tumor to brain volume ratio in TC [T2], minimum intensity in the tumor in TC [T2], intensity range within the tumor in TC [T2], sum of HOG in TC [T2], sum of entropy2 (sum entropy calculated at 0 degree and 45 degree rotations) in WT [T2-FLAIR], GLCM energy in the AT [T1-GD], median HOG in the WT [T2-FLAIR], and momentum 3rd order in the TC [T2].

The linear combination of those LASSO selected nine features enables constructing the radiomics signature. Hence, the signature score (risk score) can be calculated as follows:

$$
\begin{aligned}
Radiomics\_ & \\
signature\_score = {} & T2-FLAIR\_WT\_DifferenceVariance2\_F41 \\
& \times 0.0000018 \\
& + T2\_TC\_TumorToBrainVolumeRatio\_F79 \\
& \times 27.0110 \\
& + T2\_TC\_MinimumIntensity\_F111 \times (-0.0066) \\
& + T2\_TC\_Range\_F115 \times 0.0063 \\
& + T2\_TC\_SumHOG\_F139 \times 0.0025 \\
& + T2-FLAIR\_WT\_SumEntropy2\_F38 \times (-1.4337) \\
& + T1-GD\_AT\_Energy\_F79 \times (-3.1019) \\
& + T2-FLAIR\_WT\_MedianHOG\_F44 \times 17.1896 \\
& + T2\_TC\_Moment3rd\_F121 \times 0.0000058
\end{aligned}
$$

When the radiomics score value has been determined through the above-given signature model, the glioblastoma patient can be stratified accordingly into one of the survival groups. The thresholds, established with the ideal cutoff points on the discovery set, were low-risk (Rad-score < -0.505) for long-survivors (>15 months) group, medium-risk (Rad-score between -0.505 and 0.122) for medium-survivors (10–15 months) group, and high-risk (Rad-score > 0.122) for short-survivors (<10 months) group.

The signature model performance in both, discovery and validation, data sets stratified the patients according to the pre-determined fixed criteria/cutoff points were shown in **Figure 5**. A significant association ($P < 0.001$) of the radiomics signature

TABLE 3 | The subset of nine imaging features selected by the LASSO model and the clinical factors with their median, non-zero coefficients determined with Cox regression, and *P*-value for constructing the mpMRI radiomics signature in the discovery data set.

Characteristics	Median	Coefficients	*P*-value
Imaging features (LASSO Futures)			
T2-FLAIR_WT_DifferenceVariance2_F41	132670	1.8000e-06	9.7500e-04
T2_TC_TumorToBrainVolumeRation_F79	0.0057	27.0110	0.0028
T2_TC_MinimumTumorIntensity_F111	123.9908	−0.0066	0.0030
T2_TC_Range_F115	121.5823	0.0063	0.0030
T2_TC_SumHOG _F139	244.4848	0.0025	0.0040
T2-FLAIR_WT_ SumEntropy2_F38	1.9066	−1.4337	0.0152
T1-GD_AT_Energy_F_79	0.2027	−3.1019	0.0175
T2-FLAIR_WT_MedianHOG _F44	0.1107	17.1896	0.0185
T2_TC_Moment3rd_F121	−5324.8	5.8300e-06	0.0203
Clinical factors			
Age (years)	61.17	–	3.3700e-04
Resection status (GTR, STR, NA)	–	–	0.9720

They were ordered by their association with survival (P-value).

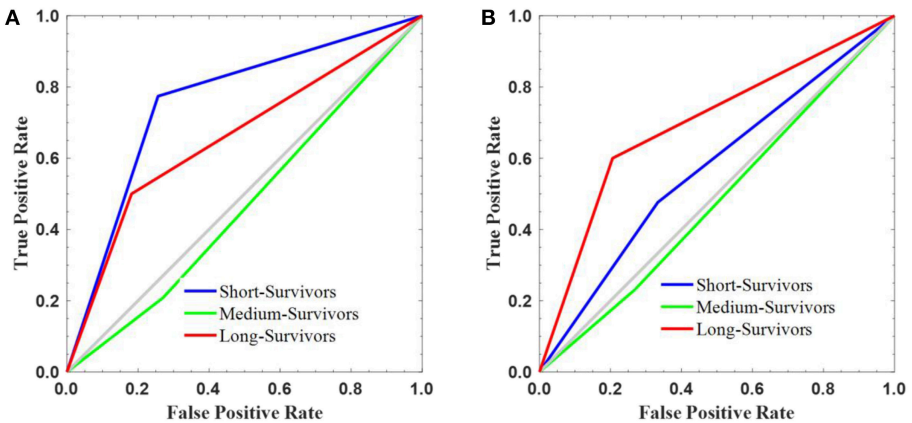


FIGURE 5 | The survival stratification created using the constructed radiomics signature. The signature performance in stratifying the survival into short-, medium-, and long-survivors on the **(A)** discovery and **(B)** validation sets.

with OS was shown in the discovery data set, but non-significant correlation ($P = 0.110$) was observed in the validation data set.

On discovery cohort, the radiomics signature stratified the GBM patients based on survival grouping with the true positive rate or sensitivity metric as following: short- (0.774), medium- (0.208), and long-survivors (0.500). The false positive rate (1—specificity) measure was 0.256, 0.271, and 0.182 for short-, medium-, and long-survivors, respectively (**Figure 5A**). In contrast, the reported values on the validation set were 0.476 (short-), 0.231 (medium-), and 0.600 (long-survivors) for true positive rate or sensitivity; and 0.333 (short-), 0.268 (medium-), and 0.206 (long-survivors) for false positive rate (1—specificity) (**Figure 5B**). For example, a false positive rate of 0.256 demonstrates that the signature model on the discovery data set assigns 26.8% of the long-survivors predictions falsely to the positive class. On the other hand, a true positive rate of 0.600 points out that the signature model classifies 60% of the predictions correctly to the positive class.

The heat map of the 9 LASSO selected features used for building the signature is shown in **Figure 6**. It shows the features association with OS between the discovery and validation data sets. From the heat map plot, it can be noticed that there is a consistency of radiomics feature z-score between the discovery/training and the validation data sets.

ML Model Results

Eight machine learning classification models were examined for survival prediction, and their performances were presented in **Table 4**. The AUC for predicting an individual survival class from the other classes, and the overall accuracy results, are reported for each model. The overall best model with feature combination for classifying OS into three groups was identified.

The best overall performance classifier was achieved by an ensemble learning model with AUC of 0.81, 0.47, and 0.72 for short-, medium-, and long-survivors (**Table 4**), respectively. The corresponding overall accuracy was 57.8% in predicting the patient's survival into short-, medium-, and long-survivors

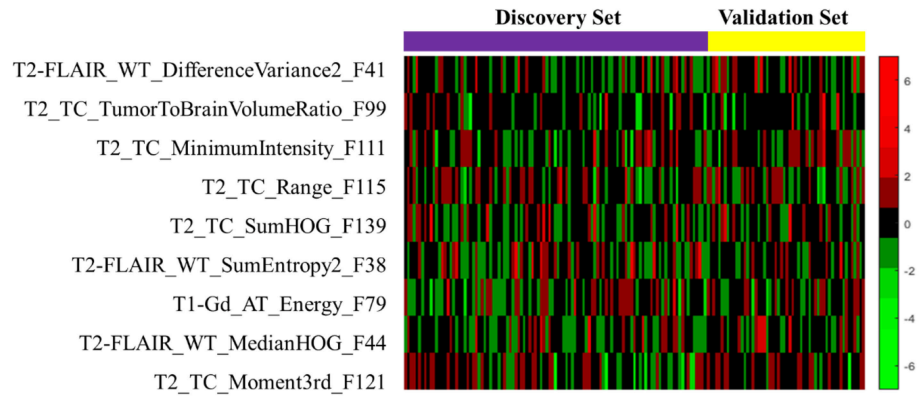


FIGURE 6 | The heat map of the LASSO selected radiomics features that used to discover the signature. The rows demonstrate the subset of nine selected features, while the columns indicate the patients (both discovery and validation data sets). The color map shows the z-score difference of each radiomics feature.

TABLE 4 | AUC and overall accuracy of several trained ML models' performance in classifying GBM patients survival into three groups as a function of choice of features.

Classifiers and features	AUC			Overall accuracy (%)
	Short-survivors	Medium-survivors	Long-survivors	
SVM (Medium Gaussian)				
• Imaging features	0.67 (0.69)	0.52 (0.60)	0.61 (0.59)	47.2 (50.3)*
• Imaging features + clinical factors	0.74	0.51	0.67	53.4
• Imaging features (LASSO)	0.72 (0.74)	0.31 (0.37)	0.68 (0.73)	50.9 (56.4)**
• Imaging features (LASSO) + clinical factors	0.80 (0.81)	0.51 (0.53)	0.68 (0.73)	54.0 (55.2)**
K-Nearest Neighbors (Coarse KNN)				
• Imaging features	0.64	0.48	0.60	46.0
• Imaging features + clinical factors	0.68	0.46	0.67	50.1
• Imaging (LASSO) features	0.73 (0.72)	0.47 (0.45)	0.72 (0.67)	47.2 (50.3)†
• Imaging (LASSO) features + clinical factors	0.79 (0.78)	0.44 (0.55)	0.70 (0.66)	47.9 (50.9)††
Discriminant analysis (Linear)				
• Imaging features	0.67	0.52	0.61	47.2
• imaging features + clinical factors	0.72	0.48	0.67	49.1
• Imaging (LASSO) features	0.74	0.45	0.72	56.4
• Imaging (LASSO) features + clinical factors	0.79	0.49	0.71	53.4
Ensemble (Random subspace discriminant)				
• Imaging (LASSO) features	0.75	0.42	0.71	57.1
• Imaging (LASSO) features + clinical factors	**0.81**	**0.47**	**0.72**	**57.8**

**Values in brackets are the performance of SVM Linear classifier.*
***Values in brackets are the performance of SVM Coarse Gaussian classifier.*
†Values in brackets are the performance of KNN Cosine classifier.
††Values in brackets are the performance of KNN Medium classifier.
The overall best classification results are listed in bold.

group. Combining the LASSO selected imaging features with the clinical predictors yielded in improved prediction accuracy results over the other alternatives in estimating glioblastoma patients' survival.

The AUC plots of the three classification models, including the ensemble model (the superior one among the other alternative models), were shown in **Figure 7**. Ideally, the perfect AUC plot is a right angle to the top left of the plot (with no misclassified points). The AUC value measures/quantifies the overall quality of the classification model. The larger AUC value demonstrates better model performance. **Figure 7** shows the AUC values for each survival class/group individually. In other words, it quantifies how the model under study is capable to classify a specific group of survivors from the other classes correctly.

Results Comparison

A comparison of this study results with other published works was presented in **Table 5**. The proposed model performance, the

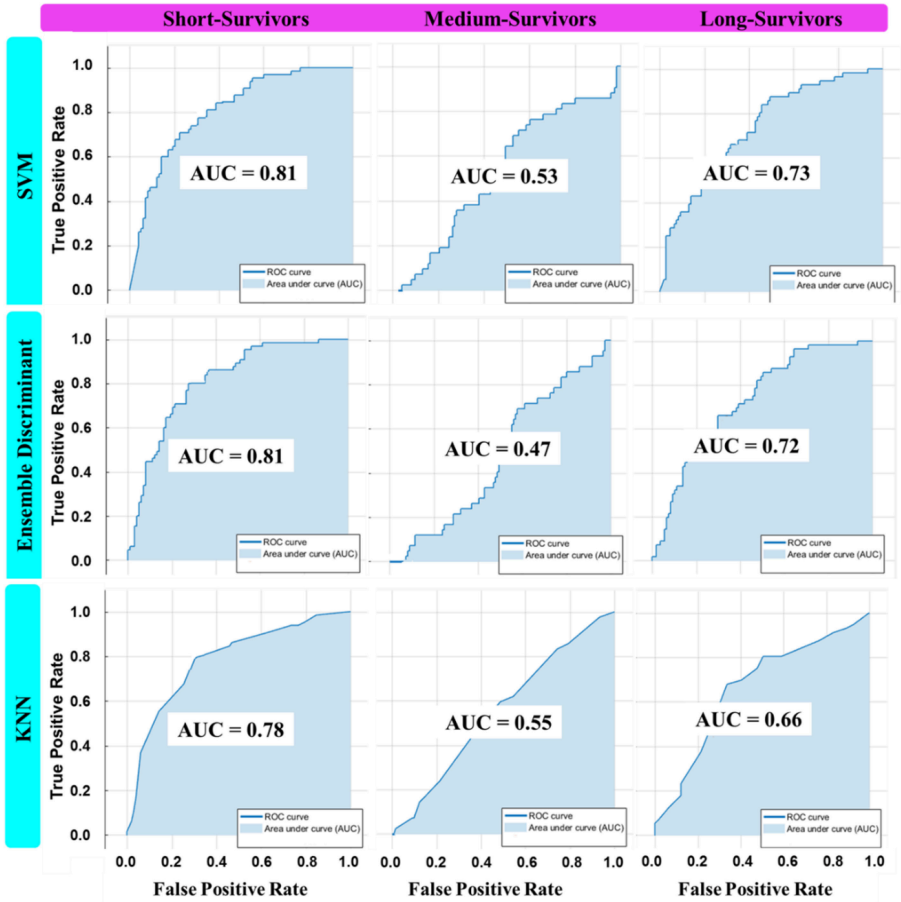


FIGURE 7 | The AUC plot of the three best overall ML classifier invariants in each machine learning category: SVM (Coarse Gaussian), KNN (Medium), and Ensemble (Subspace Discriminant) in classifying OS into three classes using the best feature combination.

signature plus the ML model, was judged amongst other works in various manners.

DISCUSSION

Radiomics analysis is the concept of extracting features quantitatively from the images/medical images using a variety of computational approaches. Then, the obtained imaging features may be used to provide clinicians with diagnosis, prognosis (e.g., survival), or treatment response. This study was aimed to identify a radiomics-based imaging signature on pre-operative mpMRI to stratify patients with *de novo* glioblastoma multiforme into short-, medium-, and long-survivors group using data from multiple institutions. Also, establishing a practical ML model for the same purpose through testing a wide range of various classification models and different features combination. Statistics, Computer Vision, and Machine Learning tools were used implementing the proposed model of radiomics analysis of patient stratification based survival grouping, which may offer unique clinical insights to support decision-making toward precision oncology.

Various image features ($n = 147$), representing tumor's shape, intensity, GLCM, HOG, and LBP (**Table 2**), were extracted and derived via different approaches on multi-parametric MRI (T1-Gd, T2, and T2-FLAIR) sequences characterizing the tumor structures [AT, TC, and WT (**Figure 1**)]. When a two-step feature selection method was employed, MAD followed by LASSO regression (**Figure 4**), a final set of 9 features retained (**Table 3**). LASSO turns all none relevant features/variables coefficients to zero during the optimization and tunes the regression model via a user-specified k–fold cross-validation. It performs both feature selection and regularization to improve the prediction accuracy and the interpretability of the statistical model it produces. The selected features indicated a high association with OS ($P < 0.05$) as shown in **Table 3**. Among those features, a gray-level co-occurrence matrix derived texture feature has shown the highest association with GBM survival stratification (**Table 3**). This finding agrees with that reported in the literature (Chaddad et al., 2015, 2016a). Image entropy and energy selected features have also shown a good correlation with survival (Chaddad et al., 2016b, 2019b; McGarry et al., 2016). Those features, typically calculated within a region of interest, indicate that intra-tumoral heterogeneity has a high impact on the survival stratification. The quantitative nature of radiomics features and the qualitative nature of radiologists to interpret the MRI sequences could

TABLE 5 | The comparison of this study's findings with similarly published works for GBM patients stratification based on survival with radiomics analysis.

Method	MRI sequences	Feature selection and classification models	Survival stratification	Overall accuracy	AUC	Signature model association with OS
Yang et al. (2015)	T1 and T2-FLAIR	Ensemble (random forest) learning	12-months survival	–	0.67	–
Macyszyn et al. (2016)	T1, T1-Gd, T2, T2-FLAIR, DTI, and DSC	SVMs	Short- (<6 months), medium- (6–18 months), and long-term (>18 months)	80.0%	–	–
This work	T1, T1-Gd, T2, and T2-FLAIR	LASSO and Cox regression, ensemble (subspace discriminant) learning	Short- (<10 months), medium- (10–15 months), and long-term (>15 months)	57.8%	0.81, 0.47, 0.72	Discovery ($P < 0.001$), validation ($P = 0.110$)
Sanghani et al. (2018)	T1, T1-Gd, T2, and T2-FLAIR	SVMs	Short- (<10 months), medium- (10–15 months), and long-term (>15 months)	88.95%	–	–
Liu et al. (2018b)	T1, T1-Gd, T2, and T2-FLAIR	SVMs	Short- (<12 months) vs. long-term (≥12 months)	80.7%	0.79	–
Chen et al. (2019)	T1-Gd	LASSO Cox regression	Short- (<12 months) vs. long-term (≥12 months)	85.1%	0.81	Discovery ($P < 0.001$), validation ($P < 0.001$)
Chaddad et al. (2019b)	T1-Gd and T2-FLAIR	Random forest	Short- (<12 months) vs. long-term (>12 months)	–	0.78	–
Zong et al. (2019)	T1, T1-Gd, T2, and T2-FLAIR	CNNs	Short- (<6 months), medium- (6–18 months), and long-term (>18 months)	64.3%,	–	–
Rathore et al. (2019)	T1, T1-Gd, T2, T2-FLAIR, DSC-MRI, and DTI	K-means clustering, Cox regression	Worst (MS = 6 months), intermediate (MS = 12 months), and longest survival (MS = 19 months)	–	–	Validation ($P < 0.001$)

DTI, Diffusion Tensor Imaging; DSC, Dynamic Susceptibility Contrast-Enhanced; CNNs, Convolutional Neural Networks; MS, Median Survival.

complementary improve the GBM patient survival prognosis quality toward precision oncology.

A multi-parametric MRI radiomics signature of 9 features was constructed on the discovery cohort for glioblastoma patients stratification based on overall survival. LASSO Cox regression model was used to extract the selected features' coefficients (**Table 3**) for developing a signature model. The author discussed the reason for applying this approach in the method section. Also, it has been reported that regression coefficients estimated by the LASSO are biased by intention, but can have smaller mean squared error than conventional estimates (Heinze et al., 2018). The radiomics signature model, trained and validated, had a good performance ($P < 0.001$) with survival association in the discovery set (n = 109), but this results not confirmed ($P = 0.110$) in the validation set ($n = 54$) (**Figure 5**). The possible reasons for non-significant results obtained in the validation set could be due to signature model overfitting during the training. It has been reported that over-fitting is possible when the number of features is greater than the number of data samples or if there are too many unique values for a discrete feature (Meinshausen and Bühlmann, 2006). The poor results obtained show the lack of generalizability of the signature model on the new unseen data set. From the statistical perspective, non-significant relationship with survival does not necessarily mean less importance (Lao et al., 2017). A second reason could be due to high contribution (almost a half, 49% as shown in **Table 1**) of patient data with missing resection status information in the combined, discovery/training and validation, cohort. These data with unknown resection information could significantly affect the overall or/and individual, training or validation, results. And, a third reason could be due to possible sub-optimal determination the cutoff points' values or thresholding in which some possibly valid assumptions had applied.

The machine learning results of several studied classifiers indicated the superiority of ensemble (Subspace Discriminant) learning over the other methods achieving the best performance accuracy of 57.8% (**Table 4** and **Figure 7**) in categorizing the survival into short-, medium-, and long-survivors. This result is not sufficiently encouraging and more tuning is needed for improved prediction accuracy. The LASSO selected imaging features, combined with clinical factors, provided better prediction results among the other options. According to the survival data distributions used in this study (**Table 1**), the best survival grouping achieved for predicting short-survivors (representing 40% of the total OS data distribution) with an AUC of 0.81. Then it followed by long-survivors (representing 36% of the total OS data distribution) with an AUC of 0.72. Finally, medium-survivors (representing 26% of the total OS

data distribution) were lasted with an AUC of 0.47. Lower performance in predicting an individual class correlated with a decreased class data distribution in the study sample. Strengths and limitations of the ML classifiers used in this study could be summarized here. Based on prediction speed, all reported models were relatively fast. In contrast, Linear models (SVM and Discriminant Analysis) are easy to interpret, SMV (with Gaussian kernels, Medium, and Coarse), KNN (Coarse, Cosine, Medium), and Ensemble (Subspace Discriminant) are hardly interpretable.

The results comparison of the proposed method (signature model and the practical ML model) with most relevant published studies are presented in **Table 5**. While the proposed method's results, the signature model and the ML model, was not impressive compared to most recently reported works (Macyszyn et al., 2016; Liu et al., 2018b; Sanghani et al., 2018; Chen et al., 2019), it was comparable or even better with respect to others studies for example that reported by Yang et al. (2015) (AUC = 0.67 for 12 months survival prediction) and Chaddad et al. (2019b) (AUC = 0.78 for short- vs. long-term OS prediction). Also, this study results are relatively comparable with that obtained by Zong et al. (2019) on multi-institutional data (accuracy of 64.3% for three-class OS prediction) using Convolutional Neural Networks, where CNN based methods are commonly expected to provide much-improved performance compared to traditional methods. The works by Macyszyn et al. (2016), and Rathore et al. (2019), reported good performance results in predicting GBM patient's survival group. However, these studies were conducted on a single institution's data, where the data is more homogeneous/consistent and more likely to obtain improved accuracy than the one used multiple institutions as a case in this study. Consequently, the model trained in local data is likely to suffer in generalizing its performance to unseen data from other institutions. On the other hand, a model trained on multi-institution data may gain generalizability but less prediction accuracy due to the heterogeneity of the data.

Finally, this study establishes that multi-parametric MR images in patients with glioblastoma hold prognostic information, which can be called up by radiomics analysis via Statistics and Machine Learning/Computer Vision methods. The proposed method in this study still has some limitations and weaknesses, which may have influenced its reported results. This work represents a retrospective study from multiple institutions with a relatively small sample patient data set used on discovery ($n = 109$) with an independent validation data set ($n = 54$) for signature model construction and evaluation. Also, almost half (49%) of the clinical data information/predictors were with no given tumor resection status (GTR or STR) information **(Table 1)**. By making available a large standard multi-institution data set, it would enable us to fully evaluate the generalizability, and thus improve the performance of the radiomics signature model on the new unseen data set.

CONCLUSIONS

Image features were extracted from pre-operative multi-parametric MR images of patients with glioblastoma to generate a radiomics signature model and a practical ML model for stratifying patients into groups based on overall survival. A derived gray-level co-occurrence matrix feature was found to have a high association with survival, which means that intra-tumoral heterogeneity has an essential role in the survival stratification. The proposed radiomics signature model had good performance in the discovery set and lower performance in the validation cohort. Despite the limitations, the offered signature model has the potential for improved pre-operative care of glioblastoma patients. Ensemble learning showed superior performance over the tested ML classifiers for survival prediction as a function of the choice of features. Clinical factors, when added to the radiomics imaging-based features, boosted the performance of the machine learning classification model in predicting individual glioblastoma patient's survival prognosis. These findings may help in choosing an optimal treatment strategy and assist in making personalized therapy decisions of glioblastoma patients which improve prognostic quality and represent a step forward toward precision oncology.

DATA AVAILABILITY

Publicly available data sets analyzed for this study. The data sets can be found in the BRATS 2018 challenge (https://www.med.upenn.edu/sbia/brats2018/data.html).

AUTHOR CONTRIBUTIONS

AO conceptualized and designed the study, developed the models, performed the statistical analysis and interpretation, created a computer graphics visualization of the results, wrote the first draft and the sections of the manuscript, and revised and approved the submitted version.

ACKNOWLEDGMENTS

The author would like to thank the BRATS'2018 Challenge team for making available large annotated multi-institution data sets that used in this study.

REFERENCES

Aum, D. J., Kim, D. H., Beaumont, T. L., Leuthardt, E. C., Dunn, G. P., and Kim, A. H. (2014). Molecular and cellular heterogeneity: the hallmark of glioblastoma. *Neurosurg. Focus* 37:E11. doi: 10.3171/2014.9.FOCUS14521

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). *Data From: Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM Collection*. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017c). *Data From: Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG Collection*. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data*. 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2019). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1811.02629v2 (accessed April 15, 2019).

Chaddad, A., Daniel, P., Desrosiers, C., Toews, M., and Abdulkarim, B. (2019b). Novel radiomic features based on joint intensity matrices for predicting glioblastoma patient survival time. *IEEE J. Biomed. Health Inform.* 23, 795–804. doi: 10.1109/JBHI.2018.2825027

Chaddad, A., Desrosiers, C., Hassan, L., and Tanougast, C. (2016a). A quantitative study of shape descriptors from glioblastoma multiforme phenotypes for predicting survival outcome. *Br. J. Radiol.* 89:20160575. doi: 10.1259/bjr.20160575

Chaddad, A., Desrosiers, C., and Toews, M. (2016b). Radiomic analysis of multi-contrast brain MRI for the prediction of survival in patients with glioblastoma multiforme. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2016, 4035–4038. doi: 10.1109/EMBC.2016.7591612

Chaddad, A., Kucharczyk, M. J., Daniel, P., Sabri, S., Jean-Claude, B. J., Niazi, T., et al. (2019a). Radiomics in glioblastoma: current status and challenges facing clinical implementation. *Front Oncol.* 9:374. doi: 10.3389/fonc.2019.00374

Chaddad, A., Sabri, S., Niazi, T., and Abdulkarim, B. (2018). Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. *Med. Biol. Eng. Comput.* 6, 2287–2300. doi: 10.1007/s11517-018-1858-4

Chaddad, A., Zinn, P. O., and Colen, R. R. (2015). "Radiomics texture feature extraction for characterizing GBM phenotypes using GLCM," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI 2015)* (New York, NY: IEEE), 84–87.

Chen, X., Fang, M., Dong, D., Liu, L., Xu, X., Wei, X., et al. (2019). Development and validation of a MRI-based radiomics prognostic classifier in patients with primary glioblastoma multiforme. *Acad. Radiol.* doi: 10.1016/j.acra.2018.12.016. [Epub ahead of print].

Collewet, G., Strzelecki, M., and Mariette, F. (2004). Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn. Reson. Imaging* 22, 81–91. doi: 10.1016/j.mri.2003.09.001

Cox, D. R., and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Dalal, N., and Triggs, B. (2005). "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (San Diego, CA: IEEE), 886–893.

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Images are more than pictures, they are data. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Haralick, R. M., Shanmugan, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 3, 610–621. doi: 10.1109/TSMC.1973.4309314

Haralick, R. M., and Shapiro, L. G. (1992). *Computer and Robot Vision*. Massachusetts: Addison-Wesley.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd Edn*. New York, NY: Springer.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. *Biom. J.* 60, 431–449. doi: 10.1002/bimj.201700067

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844. doi: 10.1109/34.709601

Holland, E. C. (2001). Progenitor cells and glioma formation. *Curr. Opin. Neurol.* 14, 683–688. doi: 10.1097/00019052-200112000-00002

Ingrisch, M., Schneider, M. J., Nörenberg, D., Negrao de Figueiredo, G., Maier-Hein, K., Suchorska, B., et al. (2017). Radiomic analysis reveals prognostic information in T1-weighted baseline magnetic resonance imaging in patients with glioblastoma. *Invest. Radiol.* 52, 360–366. doi: 10.1097/RLI.0000000000000349

Kickingereder, P., Neuberger, U., Bonekamp, D., Piechotta, P. L., Götz, M., Wick, A., et al. (2018). Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro. Oncol.* 20, 848–857. doi: 10.1093/neuonc/nox188

Lao, J., Chen, Y., Li, Z. C., Li, Q., Zhang, J., Liu, J., et al. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci. Rep.* 7:10353. doi: 10.1038/s41598-017-10649-8

Liu, X., Li, Y., Qian, Z., Sun, Z., Xu, K., Wang, K., et al. (2018a). A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. *Neuroimage Clin.* 20, 1070–1077. doi: 10.1016/j.nicl.2018.10.014

Liu, Y., Zhang, X., Feng, N., Yin, L., He, Y., Xu, X., et al. (2018b). The effect of glioblastoma heterogeneity on survival stratification: a multimodal MR imaging texture analysis. *Acta. Radiol.* 59, 1239–1246. doi: 10.1177/0284185118756951

Louis, D. N., Ohgaki, H., Wiestler, O. D., and Cavanee, W. K. (2007). *WHO Classification of Tumours of the Central Nervous System, 4th Edn*. Lyon: WHO/IARC.

Macyszyn, L., Akbari, H., Pisapia, J. M., Da X, Attiah, M., Pigrish, V., et al. (2016). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro Oncol.* 18, 417–425. doi: 10.1093/neuonc/nov127

McGarry, S. D., Hurrell, S. L., Kaczmarowski, A. L., Cochran, E. J., Connelly, J., Rand, S. D., et al. (2016). Magnetic resonance imaging-based radiomic profiles predict patient prognosis in newly diagnosed glioblastoma before therapy. *Tomography* 2, 223–228. doi: 10.18383/j.tom.2016.00250

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. New Jersey, NJ: Wiley Interscience.

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34, 1436–1462. doi: 10.1214/009053606000000281

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J.,et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Ohgaki, H., and Kleihues, P. (2005). Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. *J. Neuropathol. Exp. Neurol.* 64, 479–489. doi: 10.1093/jnen/64.6.479

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 971–987. doi: 10.1109/TPAMI.2002.1017623

Osman, A. F. I. (2018). "Automated brain tumor segmentation on magnetic resonance images and patient's overall survival prediction using support vector machines," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer), 435–49.

Patrick, E., and Fischer, F. (1970). A generalized k-nearest neighbor rule. *Inf. Control* 16, 128–152. doi: 10.1016/S0019-9958(70)90081-1

Prasanna, P., Patel, J., Partovi, S., Madabhushi, A., and Tiwari, P. (2017). Radiomic features from the peritumoral brain parenchyma on treatment-naïve multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: preliminary findings. *Eur. Radiol.* 27, 4188–4197. doi: 10.1007/s00330-016-4637-3

Rathore, S., Akbari, H., Rozycki, M., Abdullah, K. G., Nasrallah, M. P., Binder, Z. A., et al. (2019). Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Sci. Rep.* 8:5087. doi: 10.1038/s41598-018-22739-2

Sanghani, P., Ang, B. T., King, N. K. K., and Ren, H. (2018). Overall survival prediction in glioblastoma multiforme patients from volumetric, shape and texture features using machine learning. *Surg. Oncol.* 27, 709–714. doi: 10.1016/j.suronc.2018.09.002

Subramaniam, R. M. (2017). Precision medicine and PET/computed tomography: challenges and implementation. *PET Clin.* 12, 1–5. doi: 10.1016/j.cpet.2016.08.010

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Van Meir, E. G., Hadjipanayis, C. G., Norden, A. D., Shu, H. K., Wen, P. Y., and Olson, J. J. (2010). Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma. *CA Cancer J. Clin.* 60, 166–193. doi: 10.3322/caac.20069

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. New York, NY: Springer Verlag.

Yang, D., Rao, G., Martinez, J., Veeraraghavan, A., and Rao, A. (2015). Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. *Med. Phys.* 42, 6725–6735. doi: 10.1118/1.4934373

Zhang, B., Tian, J., Dong, D., Gu, D., Dong, Y., Zhang, L., et al. (2017). Radiomics features of multi-parametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clin. Cancer Res.* 23, 4259–4269. doi: 10.1158/1078-0432.CCR-16-2910

Zong, W., Lee, J., Liu, C., Snyder, J., and Wen, N. (2019). Abstract 3351: Overall survival prediction of glioblastoma patients combining clinical factors with texture features extracted from 3-D convolutional neural networks. *Proc. AACR Cancer Res.* 79(13 Suppl):3351. doi: 10.1158/1538-7445.AM2019-3351

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 20 September 2019
doi: 10.3389/fnins.2019.00966

Feature-Guided Deep Radiomics for Glioblastoma Patient Survival Prediction

Zeina A. Shboul†, Mahbubul Alam†, Lasitha Vidyaratne†, Linmin Pei†, Mohamed I. Elbakary and Khan M. Iftekharuddin**

Vision Lab in Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, United States

Glioblastoma is recognized as World Health Organization (WHO) grade IV glioma with an aggressive growth pattern. The current clinical practice in diagnosis and prognosis of Glioblastoma using MRI involves multiple steps including manual tumor sizing. Accurate identification and segmentation of multiple abnormal tissues within tumor volume in MRI is essential for precise survival prediction. Manual tumor and abnormal tissue detection and sizing are tedious, and subject to inter-observer variability. Consequently, this work proposes a fully automated MRI-based glioblastoma and abnormal tissue segmentation, and survival prediction framework. The framework includes radiomics feature-guided deep neural network methods for tumor tissue segmentation; followed by survival regression and classification using these abnormal tumor tissue segments and other relevant clinical features. The proposed multiple abnormal tumor tissue segmentation step effectively fuses feature-based and feature-guided deep radiomics information in structural MRI. The survival prediction step includes two representative survival prediction pipelines that combine different feature selection and regression approaches. The framework is evaluated using two recent widely used benchmark datasets from Brain Tumor Segmentation (BraTS) global challenges in 2017 and 2018. The best overall survival pipeline in the proposed framework achieves leave-one-out cross-validation (LOOCV) accuracy of 0.73 for training datasets and 0.68 for validation datasets, respectively. These training and validation accuracies for tumor patient survival prediction are among the highest reported in literature. Finally, a critical analysis of radiomics features and efficacy of these features in segmentation and survival prediction performance is presented as lessons learned.

Keywords: glioblastoma, segmentation, neural network, radiomics, survival prediction

OPEN ACCESS

Edited by:
Bjoern Menze,
Technical University of Munich, Germany

Reviewed by:
Suyash P. Awate,
Indian Institute of Technology Bombay, India
He Wang,
Fudan University, China

***Correspondence:**
Lasitha Vidyaratne
lvidy001@odu.edu
Khan M. Iftekharuddin
kiftekha@odu.edu

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *25 March 2019*
Accepted: *28 August 2019*
Published: *20 September 2019*

Citation:
Shboul ZA, Alam M, Vidyaratne L, Pei L, Elbakary MI and Iftekharuddin KM (2019) Feature-Guided Deep Radiomics for Glioblastoma Patient Survival Prediction. Front. Neurosci. 13:966. doi: 10.3389/fnins.2019.00966

INTRODUCTION

The World Health Organization (WHO) identifies Glioblastoma as a highly aggressive grade IV glioma. Glioblastoma is known for the presence of anaplastic glial cells along with high mitotic activity and dense cellularity, as well as the increase in microvascular proliferation (Ohgaki, 2005; Louis et al., 2007; Bleeker et al., 2012). The aggressive and infiltrative growth pattern of Glioblastoma makes curative treatment impossible, which reduces the median survival rate to less than 2-years for most patients (Johnson et al., 2013). Recently, the interest has shifted toward replacing invasive methods for tumor subtyping that predict clinical outcome with non-invasive

methods (Brown et al., 2008; Itakura et al., 2015; Yang et al., 2015). Different studies (Vartanian et al., 2014; Hu et al., 2015; Liu et al., 2017) discussed Glioblastoma heterogeneity and its implication on the clinical outcome. Glioblastoma heterogeneity can be examined through radiology images such as Magnetic Resonance Imaging (MRI) (Yang et al., 2002, 2015; Emblem et al., 2008). Quantitative radiomic imaging features (henceforth, radiomics) computed from MRI can be utilized for clinical outcome prediction (Lacroix et al., 2001; Lao et al., 2017; Shboul et al., 2017) and molecular classifications (Gutman et al., 2013; Jain et al., 2013). An accurate detection and segmentation of different abnormal tumor tissues is essential in planning treatment therapy, diagnosis, grading, and survival prediction.

Few works (Pope et al., 2005; Gutman et al., 2013; Aerts et al., 2014) have proposed different methods for predicting the survivability of patients with brain tumors. Pope et al. (2005) use different subtype tumor volumes, the extent of resection, location, size and other imaging features in order to evaluate the capability of these features to predict survival. Gutman et al. (2013) use a comprehensive visual feature set known as Visually AcceSAble Rembrandt Images (VASARI) in order to predict survival, and correlate these features for genetic alterations and molecular subtypes. Aerts et al. (2014) predict survival by quantifying a large number of radiomic image features including shape and texture in computed tomography images of lung and head-and-neck cancer patients. Several of the survival prediction studies utilize regression survival (Guinney et al., 2017; Passamonti et al., 2017) models such as the proportional hazard method while a few others utilize machine learning methods to predict survival (Macyszyn et al., 2015; Shouval et al., 2017; Kirienko et al., 2018).

Among many different feature-based and feature-learned deep neural network-based abnormal tumor tissue segmentation (Havaei et al., 2017; Mlynarski et al., 2018; Shah et al., 2018; Cheplygina et al., 2019) and survival prediction methods (Islam et al., 2013; Reza and Iftekharuddin, 2014; Vidyaratne et al., 2018) with varying performances as discussed above, there is a need to understand the effect of feature-guided deep radiomics for both tumor segmentation and patient survival prediction. A feature-guided deep radiomics approach is expected to benefit from known radiomics features that are already proven effective to guide discovery of unknown features using deep learning methods. Consequently, this work proposes a fully automated two-step survival prediction framework for patients with glioblastoma: radiomics feature-guided deep neural network methods for automated tumor tissue segmentation; and overall survival regression classification using these tumor segments and other relevant features using raw structural MRI data (Reza and Iftekharuddin, 2014; Shboul et al., 2017). The known radiomics are multiresolution fractal texture features that have shown efficacy in brain tumor segmentation (BraTS) in prior studies (Iftekharuddin et al., 2003; Islam et al., 2008; Ahmed et al., 2009; Reza and Iftekharuddin, 2014; Vidyaratne et al., 2018). The proposed framework is evaluated using two recent widely used benchmark datasets from BraTS global challenges in 2017 and 2018, respectively. Our results suggest that the proposed framework achieves better tumor segmentation and survival prediction performance compared to the state-of-the-art methods.

MATERIALS AND METHODS

The overall pipeline with each processing block used for tumor segmentation and survival prediction is shown in **Figure 1**. This fully automated method proposes a two-step survival prediction framework: radiomics feature-guided deep neural network methods for automated tumor tissue segmentation; and overall survival regression classification using these tumor segments and other relevant features. The proposed multiple abnormal tumor tissue segmentation step effectively captures both local and global feature-guided deep radiomics information in structural MRI. The survival prediction step includes two representative survival prediction pipelines that experiment with different feature selection and regression approaches.

Tumor Segmentation

The tumor segmentation methods are summarized below.

Feature-Based Brain Tumor Segmentation

This method (**Figure 2A**) utilizes several of our prior robust feature extraction algorithms to include piecewise triangular prism surface area (PTPSA) (Iftekharuddin et al., 2003), and multi-fractional Brownian motion (mBm) (Islam et al., 2008). These methods capture the non-local intensity and spatially varying texture observed in abnormal tumor tissues. In addition, several other generic features such as Texton, and raw intensity are used as input to a random forest (RF) based classifier to obtain the multi-class abnormal tumor tissue segmentation (Ahmed et al., 2009; Reza and Iftekharuddin, 2014).

Feature-Learned Brain Tumor Segmentation Using Deep CNN

This method essentially transforms the segmentation problem into an intensity-based image classification task. Localized 2D patches surrounding each pixel subjected to classification are extracted from MRI and are used as input to deep CNN architecture. We set the size of the input patch as 33×33 for tumor segmentation (Vidyaratne et al., 2018). The detailed CNN design for this method is shown in **Figure 2B**.

Feature-Learned Brain Tumor Segmentation Using Deep U-Net

This method utilizes a CNN based U-Net model (Ronneberger et al., 2015; Dong et al., 2017) to obtain brain tumor segmentation. U-Net model is known for end-to-end data processing. Unlike patch based CNN segmentation pipeline where the model only sees a localized region of the brain, the U-Net in this work captures global information from different regions of the brain, which is essential to achieve robust segmentation performance. The U-Net architecture utilized in this work is implemented following the work in Dong et al. (2017). More specifically, the architecture consists of a down-sampling (encoding) and an up-sampling (decoding) stage. The down-sampling stage has five convolutional blocks each consisting of two convolutional layers with a filter size of 3×3

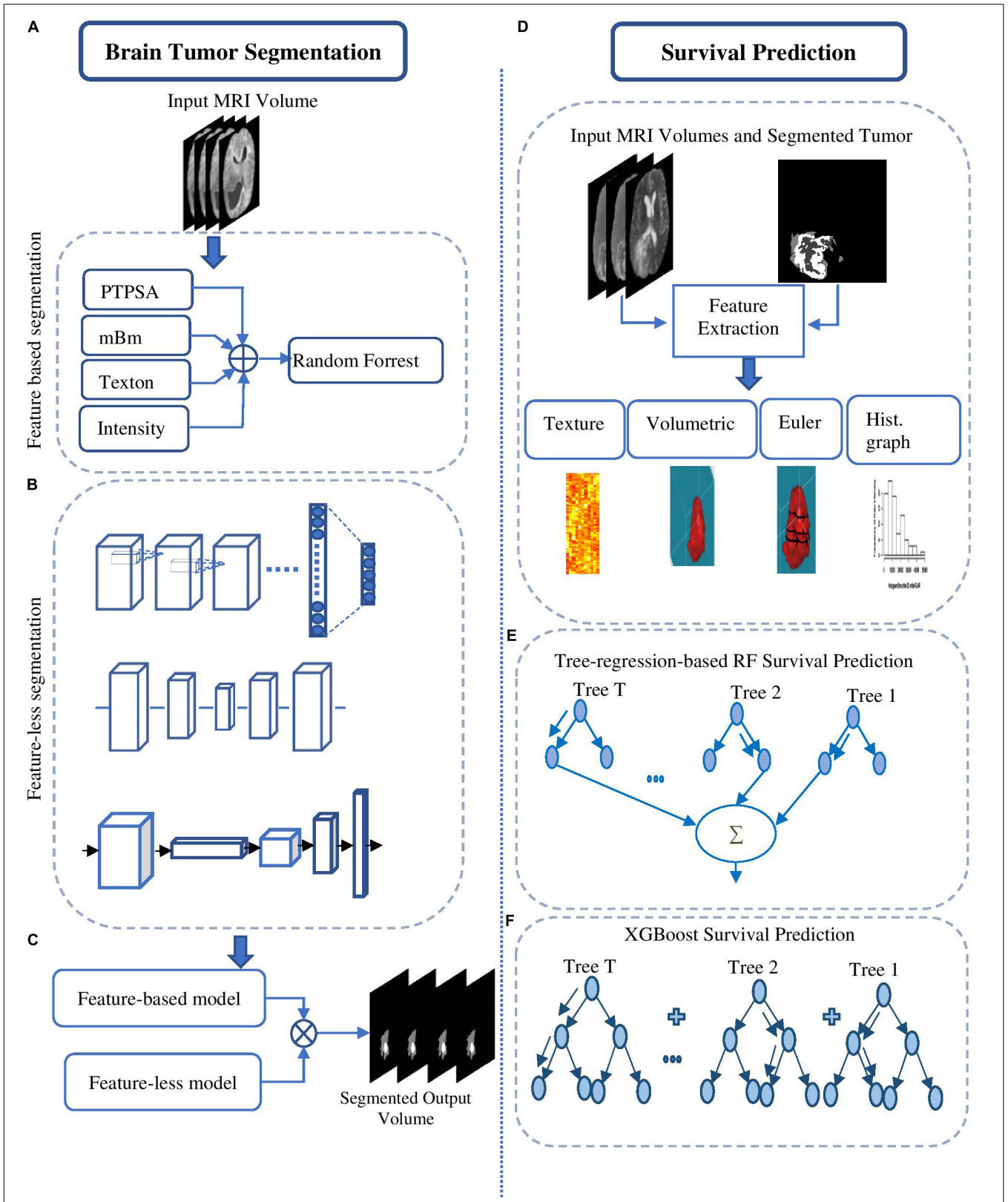


FIGURE 1 | Brain tumor segmentation pipelines (left) using: **(A)** feature-based high-grade tumor segmentation using RF; **(B)** feature-less segmentation using Deep CNN, U-Net, and FCN; and **(C)** semantic label fusion using feature-less and feature-based. Survival prediction pipelines (right) are started with **(D)** feature extraction and are trained using **(E)** tree-regression-based RF survival prediction, and **(F)** XGBoost-based survival prediction.

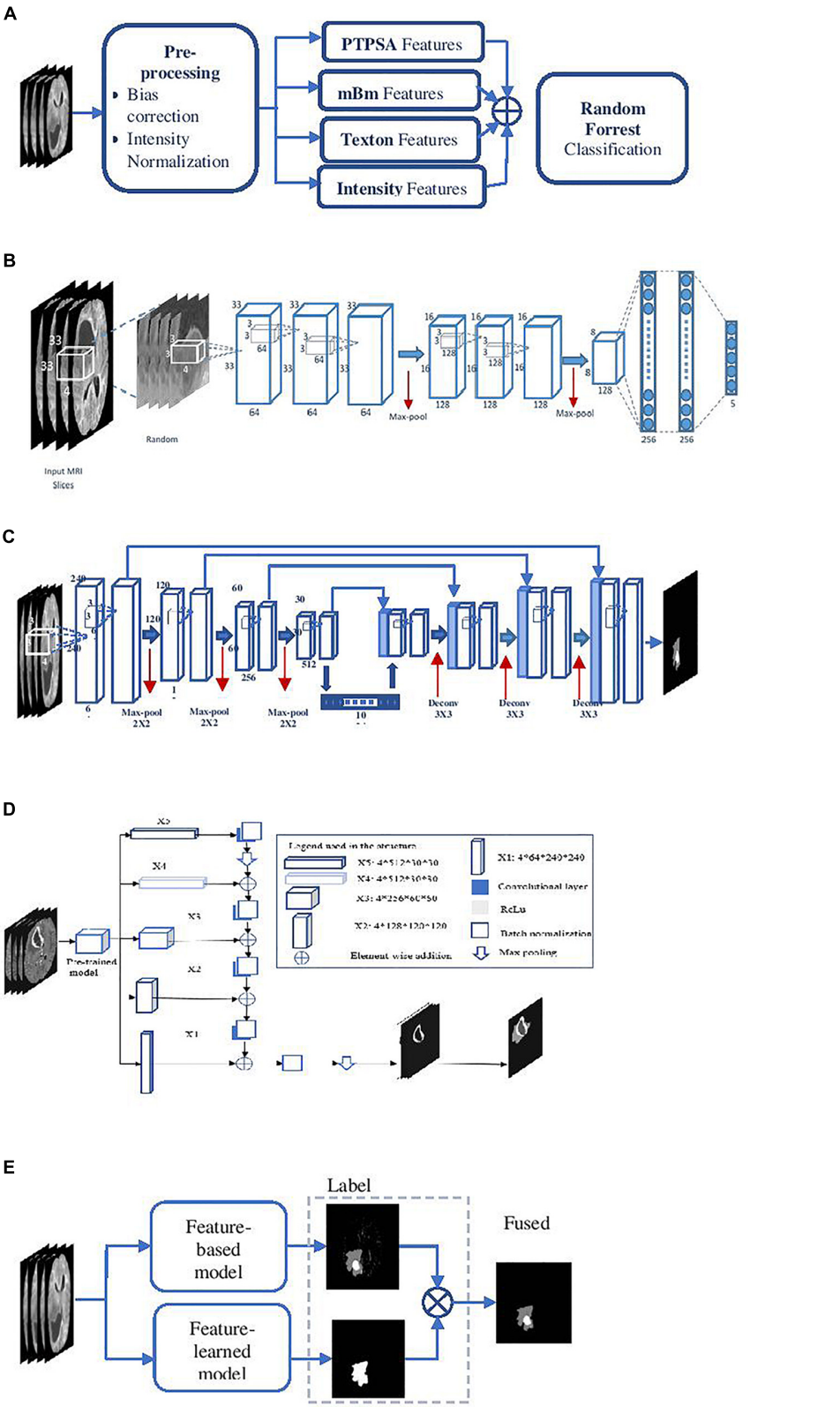


FIGURE 2 | Overall segmentation pipelines used in the proposed methods. **(A)** Feature-based High-grade tumor segmentation using RF; **(B)** detailed architecture of the CNN based high-grade tumor segmentation; **(C)** Low-grade tumor segmentation with U-Net (detailed architecture); **(D)** architecture of brain tumor segmentation (BraTS) using FCN; and **(E)** general pipeline of BraTS fusion by feature-based and feature-learned model.

and stride of 1 followed by maxpooling with stride 2 × 2. The upsampling stage consists of deconvolution layer with a filter size of 3 × 3 and stride of 2 × 2 which doubles the size of the feature maps. Rather than using regular cross-entropy based loss function, we utilize a soft dice metric based loss function to train the U-Net model (Milletari et al., 2016). The soft dice is a differentiable form of the original dice similarity coefficient (DSC) which is the most widely used metric to evaluate tumor segmentation performance. The model is trained using mini-batch gradient descent (GD) technique which minimizes the soft dice cost function. **Figure 2C** shows the detailed architecture of the U-Net model to perform the BraTS task.

Feature-Learned Brain Tumor Segmentation Using Fully Convolutional Networks

Fully convolutional networks (FCNs) have been successfully used for many image processing and computer vision tasks (Long et al., 2015; Zhao et al., 2016). FCNs build FCNs that take an input of arbitrary size and produce a correspondingly sized output of relevant characteristics with efficient inference and learning. Accordingly, FCN contains only convolutional layers. It removes any redundancy when computing classification maps on large inputs. The architecture also features an encode (down-sampling) and a decode (up-sampling) stage. The encode stage of the proposed architecture has five convolutional blocks. Each block is composed of two convolutional layers with a filter size of 3 × 3 and stride of 1 followed by maxpooling with stride 2 × 2. The decode stage consists of deconvolution layers with a filter size of 3 × 3 and stride of 2 × 2 which doubles the size of the feature maps. The framework of the proposed method is shown in **Figure 2D**, which uses VGG-11 (Simonyan and Zisserman, 2014) as a pre-trained model.

Semantic Label Fusion of Feature-Based and Feature-Learned Deep Radiomics for Improved Tumor Segmentation

The different deep radiomics-based models discussed above are first independently implemented and trained for multi-class abnormal tumor tissue segmentation. In order to complement both feature-based and feature-learned radiomics methods, we implement a label fusion method (**Figure 2E**) for improved tumor segmentation. The label fusion is then performed to obtain the fused output F_i^v for volume v as follows:

$$F_i^v = U_i \bigcup_{i \in v} C_i; \tag{1}$$

WhereU_i, and C_i denote the U-Net and FCN outputs given MRI volume v, respectively.

The outputs of U-Net and FCN architectures offer excellent specificity, albeit with varying sensitivity performance. The union operation in equation (1) essentially preserves the specificity while improving the sensitivity by combining the within-class regions from each output. Similarly, this method is used for label fusion between the patch-wise CNN based segmentation algorithm and the hand-crafted feature-based algorithm for better segmentation performance.

Survival Prediction

The survival prediction model includes prediction of survival risk classification (short, medium, and long-term survival). Subsequently, an overall survival regression is performed based on the survival risk class label. Both classification and regression models are trained on quantitative- radiomics features obtained from the segmented tumor. Recursive feature selection (RFS) method is used to select the features that are used in the classification model. Finally, Cox regression is used as a feature selection method in the overall survival regression model. Three overall regression models are trained: long-regression model, mid-regression model, and short regression model.

Feature Extraction

Feature extraction is the first step of the overall survival prediction task. Different quantitative imaging features (of around 31,000) are extracted from the different types of segmented abnormal tissues (edema, enhancing tumor, and tumor core) obtained in the previous step. These features include texture, volumetric and area-related features, histogram-graph features, and Euler characteristics (vertices, edges, and faces). The heterogeneity in Glioblastoma may be quantified using texture and histogram-graph features; while the shape of the tumor may be effectively captured using volumetric and Euler characteristic features (Pope et al., 2005; Aerts et al., 2014; Rathore et al., 2016).

A detailed breakdown of the extracted features is as follows: a total of 1107 texture features (Vallières et al., 2015) are computed from raw MRI sequences, and the features are extracted from eight texture representations of the tumor volume [Texton filters (Leung and Malik, 2001); texture-fractal characterization using both our PTPSA (Iftekharuddin et al., 2003) modeling and multi-resolution mBm (Islam et al., 2008) modeling; and the characterization Holder Exponent (Ayache and Véhel, 2004) modeling of the tumor region]. Furthermore, six histogram-based statistics (mean, variance, skewness, kurtosis, energy, and entropy) features are extracted from the edema, enhancing tumor, and necrosis tissues.

Moreover, 13 volume-related features are considered: the volume of the whole tumor, the volume of the whole tumor with respect to the brain, the volume of sub-regions (edema, enhancing tumor, and necrosis) divided by the whole tumor, the volume of sub-regions (edema, enhancing tumor, and necrosis) divided by the brain, the volumes of the enhancing tumor and necrosis divided by the edema, the summation of the volume of the edema and enhancing tumor, the volume of the edema divided by the summation of the volume of enhancing tumor and necrosis, and the volume of the necrosis divided by the summation of the volume of the edema and enhancing tumor. The tumor locations and the spread of the tumor in the brain are computed. Another nine area-related properties (area, centroid, perimeter, major axis length, minor axis length, eccentricity, orientation, solidity, and extent) are computed from three viewpoints (x, y, and z-axes) of the whole tumor.

Furthermore, a total of 832 features are extracted from the histogram graph of the different modalities of the whole tumor, edema, enhancing and necrosis regions. These features represent the frequency at different intensity bins (of 11,15, and 23) and

the bins of the max frequency. Finally, we compute the Euler characteristic (Turner et al., 2014) of the whole tumor, edema, enhancing and necrosis, for each slice. The Euler characteristic features are computed on the tumor curve, at 100 points, and at 72 different angles. Then, the Euler characteristic features are integrated over all the slices. As a result, each patient is represented by 4 (whole tumor, edema, enhancing, and necrosis) Euler characteristic feature vectors. Each vector has a size of 7200 (100 points $\times$ 72 angles).

Survival Prediction Models

Two different survival prediction models are proposed for survival prediction. The first model is a tree-based method for overall-survival regression prediction using RF regression model. We have employed RF due to its efficiency, robustness and the flexibility in utilization for both multi-class classification and regression tasks (Breiman, 2001). Additionally, RF does not require extensive hyper-parameter tuning, and is resilient to overfitting. These traits make RF preferable over more common models such as artificial neural networks especially when the training data is limited. The complete pipeline for the survival regression using RF is illustrated in **Figure 3A**. This model uses significant, predictive and important features selected from the above-mentioned texture, histogram-graph, and volumetric and area-related features. A three-step feature selection method is utilized as follows. A univariate cox regression is fitted on every extracted feature, and features with p-value less than 0.05 are considered as significant. A second univariate cox regression is fitted on the quantitative copy of the significant features. The quantitative copy is obtained by thresholding the significant feature around its median value. The last step is performed to ensure that each significant feature is also able to split the data set into long vs. short survival. Then, RF regression model with tenfold cross validation is used to evaluate the model at each iteration.

The model in **Figure 3A** is used as a baseline to obtain a second more comprehensive survival prediction pipeline as shown in **Figure 3B**. We incorporate additional features such as Euler characteristics. The features for the updated model are then selected using RFS method as follows. First, we perform RFS1 on the Euler features alone. Next, another RFS2 on the remaining features (texture, volumetric, histogram-graph based) is performed. In addition, the overall-survival regression model uses Cox regression to select significant features with p-value < 0.05. Moreover, we introduce a state-of-the-art Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) based regression technique for stepwise survival risk classification and overall-survival regression prediction using the selected features. The XGBoost based regression model is applied to each of the three groups (short, medium, and long) to obtain survival duration in the number of days, respectively. One of the major advantages of XGBoost its utilization of L1 and L2 regularization. L1 regularization handles sparsity, whereas L2 regularization reduces overfitting (Chen and Guestrin, 2016).

It is worth noting that we have not utilized any neural network model for the survival prediction because the sample size in this study is not large enough to ensure good training in a neural network setting.

RESULTS

Dataset

This study uses BraTS18 training, validation and testing dataset (Menze et al., 2015; Bakas et al., 2017a,b), and BraTS17 training, validation, and testing datasets for patient survival prediction analysis. Both BraTS17 and BraTS18 datasets contain a total of 163 Glioblastoma [high grade glioma (HGG)] cases for training, with an overall survival, defined in days, and the age of patient at diagnosis, defined in years. The training dataset provides four modalities [T1, post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR)] along with the ground truth segmentation of multiple abnormal tissues (enhancing, edema, necrosis, and non-enhancing) in the tumor. Overall survival risk is classified into three survival groups: long (greater than 15 months), medium (between 10 and 15 months), and short (less than 10 months). In addition, for validation purposes, we use the validation datasets of BraTS17 and BraTS18. BraTS17 validation dataset consists of 33 cases while that for BraTS18 consists of 28 cases for overall survival prediction purposes. BraTS17 testing dataset consists of 95 cases while that for BraTS18 offers 77 cases for testing the overall survival prediction performance.

Overall Survival Prediction Framework Evaluation

As discussed in the Methods section, the proposed framework consists of several feature-based and feature-guided deep radiomics-based automated BraTS methods and two distinct deep radiomics based automated survival prediction pipelines. Accordingly, we obtain extensive performance evaluation using two pipelines: the first one combines CNN-based patch-wise segmentation algorithm, radiomics feature-based segmentation algorithm, and RF based survival prediction method (henceforth SP1), while the second combines U-Net and FCN based segmentation methods with the XGBoost based survival prediction algorithm (henceforth SP2). We first participated in the BraTS 2017 challenge and the specific combination of machine learning methods with RF survival prediction model (known as SP1) offered the best overall performance in this Challenge. We subsequently participated in the BraTS 2018 challenge and the augmented model (known as SP2) offered the best performance using the validation dataset. The mean dice segmentation performance (of enhancing tumor, whole tumor, and tumor core) for SP1 and SP2 is illustrated in **Table 1**. The mean dice segmentation metrics for different sub-tissues are evaluated using the online evaluation platform of the BraTS challenge (CBICA IPP at[1]). A detailed performance analysis of U-Net, FCN and their sematic-label fusion results are illustrated in **Table 2**. **Figure 4** shows an example of segmentation outcomes using U-Net, FCN and semantic-label fusion of U-Net and FCN.

[1]https://ipp.cbica.upenn.edu

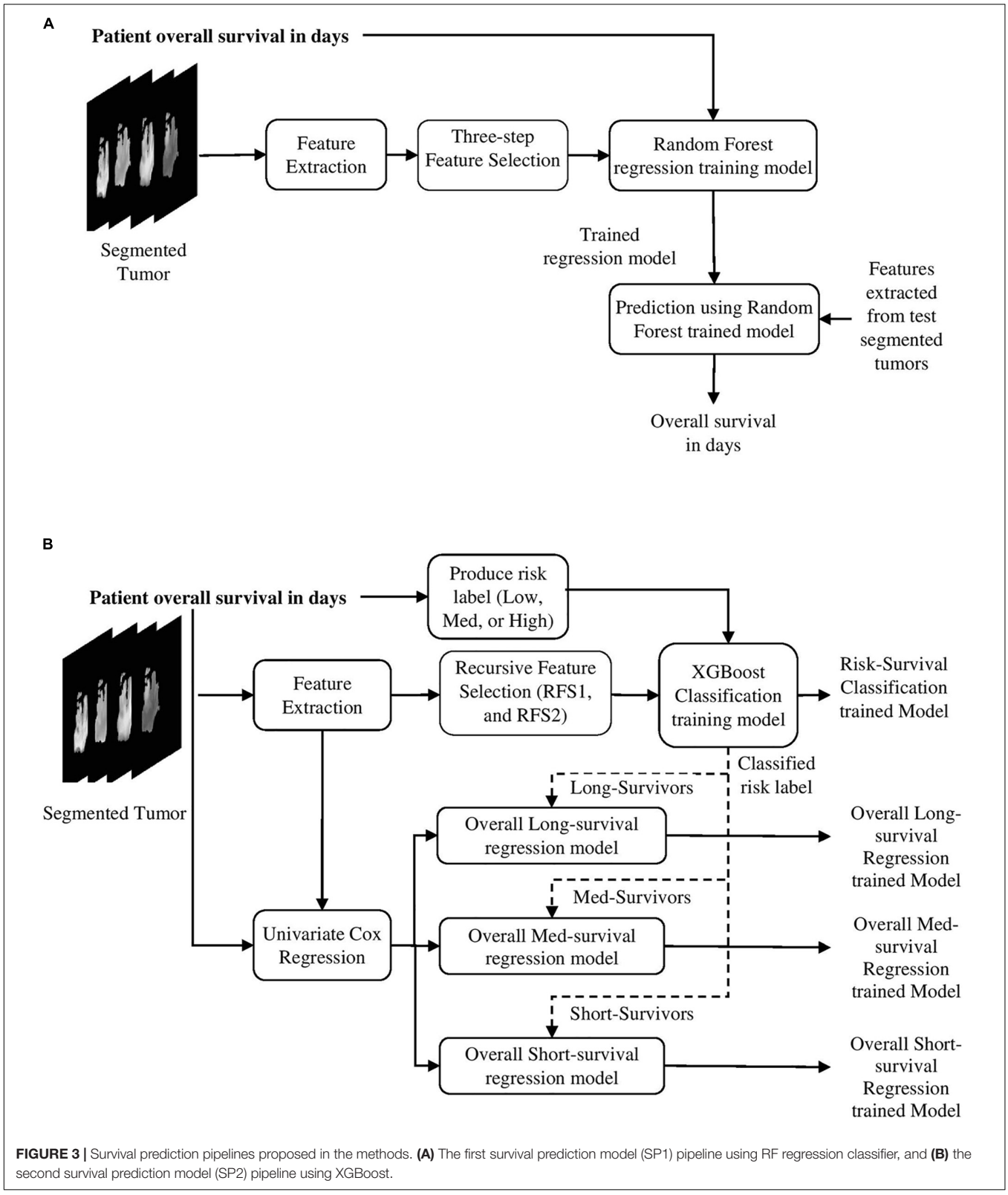


FIGURE 3 | Survival prediction pipelines proposed in the methods. **(A)** The first survival prediction model (SP1) pipeline using RF regression classifier, and **(B)** the second survival prediction model (SP2) pipeline using XGBoost.

For SP1 the survival prediction features are the age and 40 texture and volumetric features. The distribution of the 40 features is as follows: 12 features extracted from Texton of the tumor, 9 features extracted from the Holder exponent representations of the tumor, 6 features represent the histogram of the abnormal tissues, 5 from the raw MR modality of the tumor

TABLE 1 | Performance of SP1, SP2, and modified-SP2 methods with BratS17 and BraTS18 datasets.

	Survival prediction performance		Segmentation performance		
Model/dataset	**Accuracy**	**MSE**	**Dice enhanced tumor**	**Dice whole tumor**	**Dice tumor core**
SP1/BraTS17 training	0.67	78,929	–	–	–
SP1/BraTS17 validation	0.667	2,09,908	0.746	0.815	0.698
SP1/BraTS17 test	0.579	2,45,780	0.733	0.832	0.725
SP2/BraTS18 training	0.73	91,585	–	–	–
SP2/BraTS18 validation	0.679	1,53,466	0.765	0.876	0.761
SP2/BraTS18 test	0.519	3,67,240	0.705	0.857	0.767
RF-SP1/BraTS18 validation	0.464	1,70,737	–	–	–
XGBoost-SP2/BraTS17 validation	0.636	2,18,097	–	–	–
Modified-SP2/BraTS18 training	0.718	99,358	–	–	–
Modified-SP2/BraTS18 validation	0.679	1,27,697	–	–	–

The evaluation of validation is performed using the online evaluation platform of CBICA IPP (https://ipp.cbica.upenn.edu).

TABLE 2 | Performance of U-Net, FCN and their Semantic-label fusion using BraTS18 validation dataset.

Model	**Dice enhanced tumor**	**Dice whole tumor**	**Dice tumor core**
FCN	0.706	0.850	0.727
U-Net	0.697	0.835	0.719
Semantic-label fusion	0.714	0.861	0.740

and sub-regions, 4 describe the volume of the tumor and the sub-regions, and 4 features are extracted from the tumor area and major axis length.

In comparison, as discussed above and shown in **Figure 3B** for SP2, all relevant features are extracted from the ground truth cases available with BraTS18 training dataset. The subsequent RFS for Euler features (28,000) alone generates 39 features. The distribution of the 39 Euler features includes: 16 features computed around the contour of ET, 16 features computed around that of WT, and 7 features computed around that of edema, respectively. The application of RFS on the remaining features produces additional 23 texture features, 4 histogram graph features, and 8 area features of the edema, ET, and WT, respectively. The XGBoost with leave-one-out cross-validation (LOOCV) is employed on the selected 74 features and the age to predict three corresponding survival classes (short, medium, and long). This yields a classification accuracy of 0.73 [95% confidence intervals (CI): 0.655–0.797] for the BraTS18 training dataset.

First, we establish the performance of both SP1 and SP2 methods using the BraTS17 and BraTS18 training, and validation datasets. The training dataset performance is obtained through LOOCV analysis. The performance evaluation of methods using BraTS validation datasets is restricted to the online evaluation platform of the organizer of the BraTS challenge and must be performed during a specific time period during the challenge. Note that the second pipeline (SP2) is developed after the BraTS 2017 challenge is concluded, and hence 2017 validation portal is no longer available for evaluation. However, a fair comparison between the pipelines can still be obtained through the training data evaluations and the validation evaluations of respective challenge years. The results are summarized in **Tables 1**, **3**.

The results in **Table 1** for training and validation illustrate that SP2 model offers better performance in accuracy over that of SP1 model. SP2 model also obtains improvement over SP1 in validation MSE. This performance improvement may be attributed to improved abnormal tumor tissue segmentation as well as the use of additional features obtained using better feature selection and regression methods. Note that SP1 model has been ranked the first in the BraTS 2017 challenge for survival prediction category among 17 teams globally. The overall high MSE for survival prediction is particularly due to the wide range within long term survival category resulting in large prediction errors. Further, note that the MSE of SP2 for the BraTS18 training is the sum of the three MSE (**Table 4**) values obtained for the short-, medium-, and long-regression models shown in **Table 4**. Finally, the test results for both SP1 for BraTS17 and SP2 for BraTS18 in **Table 1** show that SP1 performed better in patient-survival prediction than that for SP2. This performance difference for SP1 and SP2 models is further analyzed below.

Comparative Evaluation of Survival Prediction Performance With SP1 and SP2

Table 3 shows the confusion matrix of both SP1 and SP2 and relevant statistics for each class in the classification training model for survival risk prediction. The sensitivity and balanced accuracy of the medium survival group in SP2 is the lowest when compared to the other two survival groups.

The top four important features as ranked by XGBoost are: tumor extent in z-axis, the width of the enhance tumor computed from x-axis point of view, contour around the edema contour and enhance tumor. The mean value of each of these four features is able to significantly (p-value < 0.05) stratify the 163 cases into two risk groups (low-risk and high-risk) as illustrated in **Figure 5**.

The second step in the survival prediction is to obtain individual regression training models corresponding to the short, medium, and long survival classes. These short-, medium-, and long-regression models use features selected distinctly for each survival class using Cox regression (with p-value < 0.05). The number of significant features selected for the short-, medium-, and long-regression models are 83, 51, and 148, respectively.

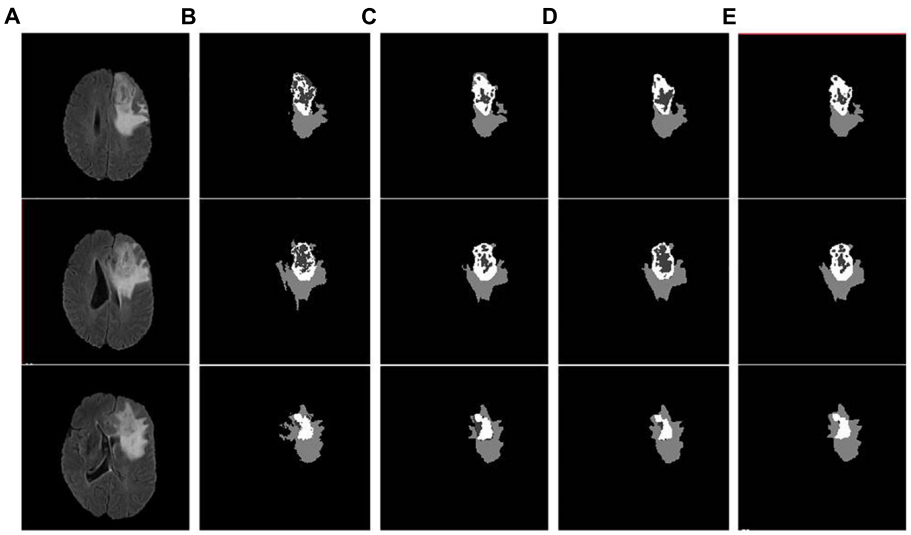


FIGURE 4 | Example input slices from BraTS18 training dataset and segmentation outcomes: **(A)** Flair sequence; **(B)** the ground truth; **(C)** the segmentation outcome of U-Net; **(D)** the segmentation outcome of FCN; and **(E)** semantic label fused segmentation.

TABLE 3 | Confusion matrix of SP1, SP2, and modified-SP2, and some statistics derived from the confusion matrix based on each survival label in the training model.

	SP1 2017			SP2 2018			Modified-SP2 2018		
	Reference			Reference			Reference		
	Long	Med	Low	Long	Med	Low	Long	Med	Low
Predictions									
Long	32	7	10	43	13	4	44	11	4
Med	24	34	12	5	18	3	7	18	6
Low	0	1	43	8	11	58	5	13	55
Total number of cases	56	42	65	56	42	65	56	42	65
Statistics									
Sensitivity	0.571	0.810	0.662	0.768	0.429	0.892	0.786	0.429	0.846
Specificity	0.841	0.702	0.990	0.841	0.934	0.806	0.860	0.886	0.816
Balanced accuracy (Sen + Spec)/2	0.706	0.756	0.826	0.804	0.681	0.849	0.823	0.657	0.831
Positive prediction value (PPV)	0.653	0.486	0.977	0.717	0.692	0.753	0.745	0.581	0.753
Negative prediction value (NPV)	0.789	0.914	0.815	0.874	0.825	0.919	0.885	0.817	0.889

TABLE 4 | Performance of LOOCV of the three regression models in SP2 and modified-SP2 in the XGBoost overall survival model.

	SP2			Modified-SP2		
	Root mean square error (RMSE)	MSE	Mean absolute error (MAE)	Root mean square error (RMSE)	MSE	Mean absolute error (MAE)
Long-regression model	294.177	86,540	217.714	302.069	91,246	209.253
Medium-regression model	35.629	1,269	28.190	40.702	1,657	34.971
Short-regression model	61.449	3,776	50.402	80.340	6,455	65.094

Table 4 illustrates the performance of LOOCV with XGBoost for the selected features using specified survival risk cases in BraTS18 training cases.

Note that the wide range of the overall survival of the long-survival group (greater than 15 months) may cause the RMSE of the long-regression model to have the highest RMSE (**Table 4**). This also may cause the high mean square error when using the validation dataset (**Table 1**). The range of the overall survival of the short-survival group is 10 months, whereas the medium-survival group is 5 months.

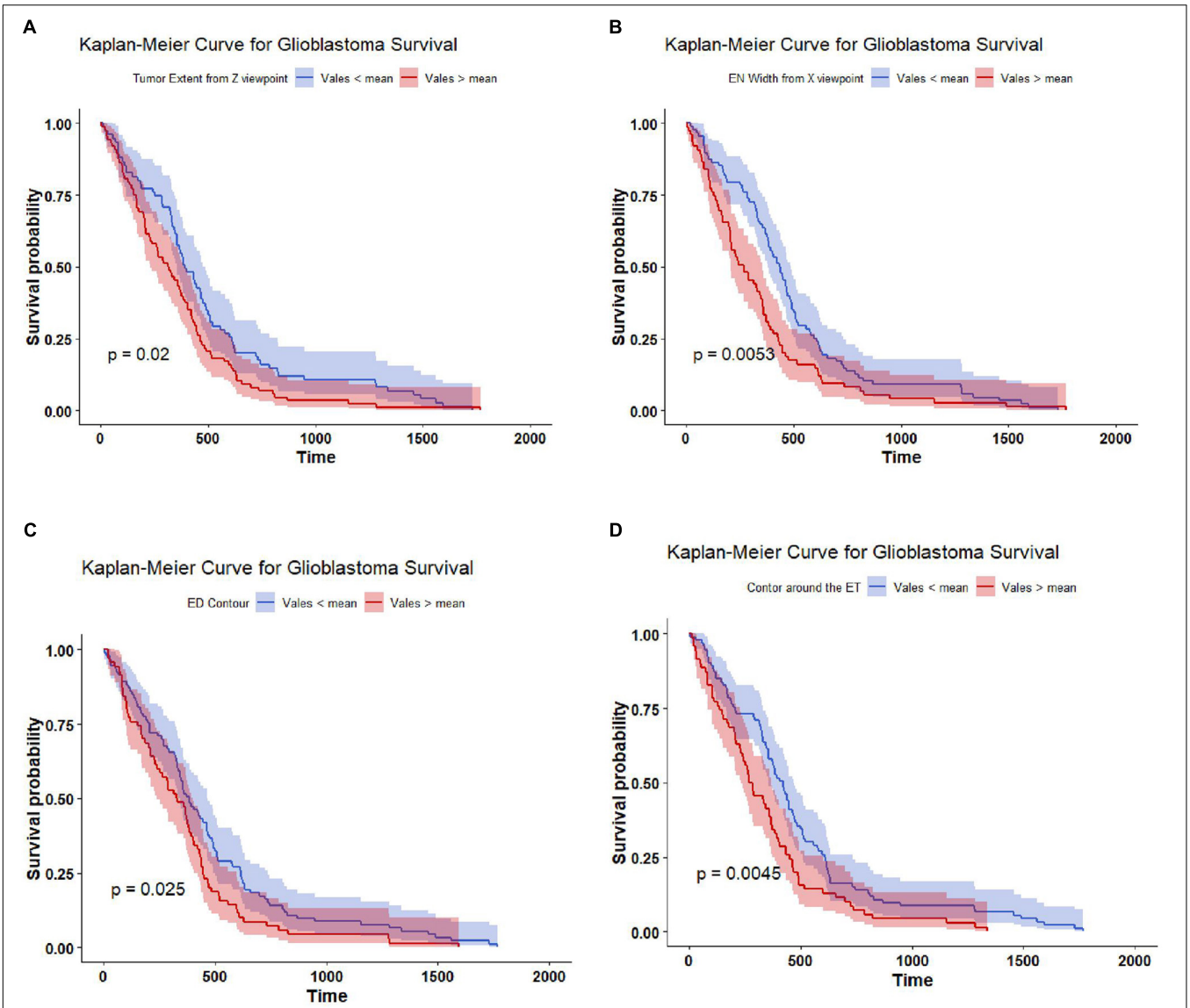


FIGURE 5 | Kaplan Meier of the top four important features used in SP2. The features are thresholded around its mean value to stratify the 163 subjects into two groups: high risk group (red line), and low risk group (blue line). The features are **(A)** tumor extent; **(B)** enhance tumor width; **(C)** contour around the edema; and **(D)** contour around the enhance tumor. The shaded area indicates the 95% confidence interval.

Critical Analysis of Features and Performance of the Survival Prediction Pipelines

This section provides a critical analysis of the features and their effect on the survival prediction performance. As mentioned in the previous sections, the features that are derived from different abnormal tissue types of the segmented tumor region significantly contribute to the survival prediction performance (the abnormal tissue segmentation dice performance of SP1 and SP2 are illustrated in **Table 1**). Accordingly, we visualize the features extracted from different abnormal tissue types of the segmented tumor. The visualization is performed using one of the most widely used high-dimensional data visualization techniques known as t-Distributed Stochastic Neighbor Embedding (Maaten and Hinton, 2008) (t-SNE). First, t-SNE is used to explore the features obtained from different abnormal tissue types from the segmented tumor region and analyze the effect of these features on the performance of the survival prediction task using BRAST 2017 and BRAST 2018 dataset.

For the SP1 pipeline, we extract a total of 40 features from the sub-tissue types of the segmented tumor region. The features extracted in SP1 are as follows: 36 features for whole tumor, 2 features for enhanced tumor, and 2 features for edema. **Figures 6A–C** shows a visualization of these features across different abnormal tissue types for BraTS17 training, validation and testing data, respectively. These figures demonstrate that the extracted features for segmentation offer clear discrimination

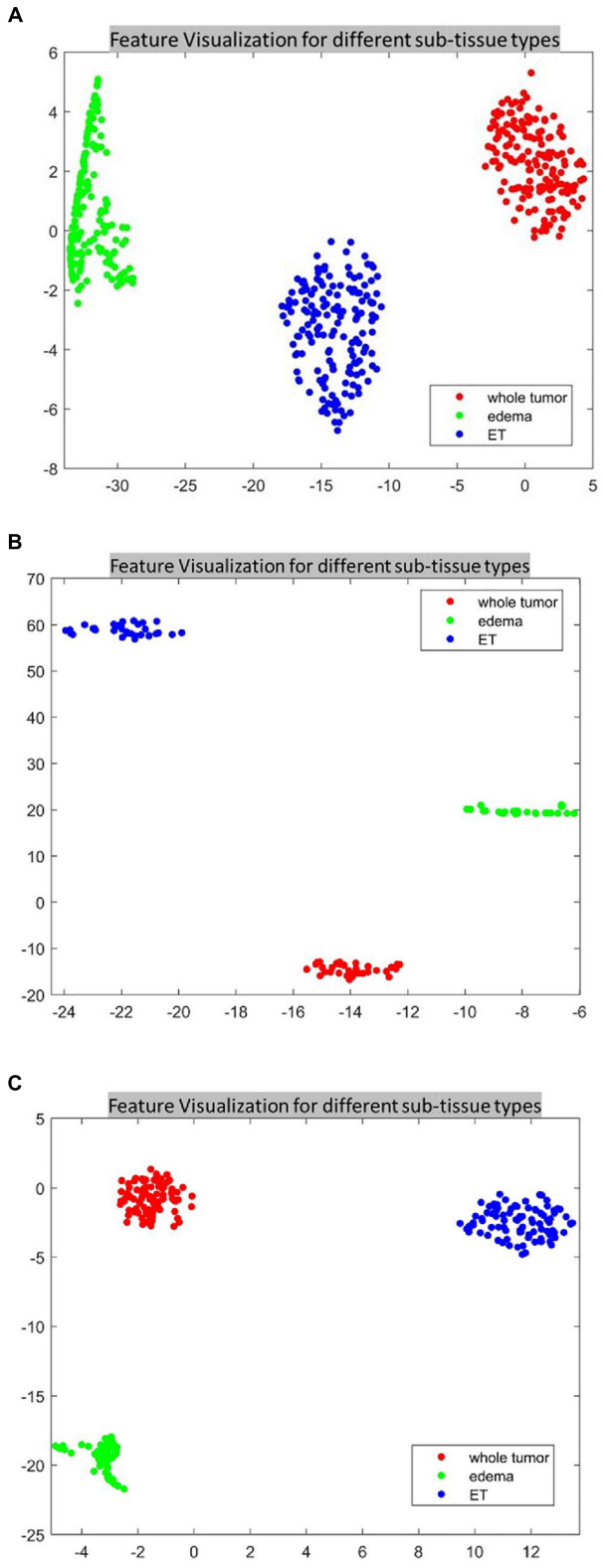


FIGURE 6 | The t-distributed stochastic neighbor embedding (t-SNE) of the selected features of SP1 clustered based on their tissue types using BraTS17 **(A)** training; **(B)** validation; and **(C)** testing. Note that features are clustered based on their origin (subtissue type).

among different abnormal tissue types in the tumor. This demonstrates the effectiveness of the segmentation pipeline in SP1. Next, we visualize the feature clusters for patient survival categories: long, medium and short term. In this case we consider all 40 features obtained from the 163 BraTS17 training data as mentioned above and explore the grouping against the tumor risk labels using the t-SNE technique. **Figure 7** shows the visualization of the corresponding features for long, medium and short risk labels. Note that all the visualization outcomes shown are obtained after extensive hyper-parameter tuning of t-SNE to produce the best possible results. **Figure 7** demonstrates that though there is some separation of corresponding features between the long and short categories, the medium category is mixed with both long and short categories. This suggests that it is still difficult to visualize a clear separation of extracted features for survival prediction task with the available patient dataset for this study. The corresponding survival prediction performance of SP1 pipeline using testing dataset is as shown in **Tables 1**, **3**. As mentioned above, though the SP1 pipeline was ranked the first place in BraTS 2017 challenge, the feature distribution in **Figure 7** suggests inherent challenge in extracting representative features for survival prediction task.

Next, we explore the features and their effect on the performance of our SP2 pipeline using the BraST18 dataset. We extract a total of 74 features and the age for the SP2 pipeline. The features extracted in SP2 are as follows: 43 features for whole tumor, 22 features for enhanced tumor, and 8 features for edema, and 1 feature for necrosis. **Figures 8A–C** shows a visualization of these features across different tissue types for BraTS18 training, validation and testing data, respectively. **Figure 8** demonstrates that these features also offer a clear separation for different abnormal tissue types in the tumor. Therefore, this further demonstrates the effectiveness of our segmentation pipeline in SP2 and verifies that the extracted features are highly representative of the different abnormal tissue regions (the abnormal tissue segmentation dice performance of SP2 is illustrated in **Table 1**). Subsequently, **Figure 9** shows the visualization of the 74 features in terms of long, medium and short risk labels using the 163 sample BraTS18 training data. Our analysis suggests that the tSNE technique again fail to group the features in long, medium and short categories. Though there is some separation between the corresponding features for long and short categories, the features for medium category mixes with both short and long categories for multiple subjects, quite similarly to the visualization of SP1. This poor separation may still be due to the lack of sufficient representative strength of the features for categorizing different risk labels. Consequently, **Table 1** shows that our proposed SP2 pipeline achieves 0.73, 0.679, and 0.519 accuracy on the BraTS18 training, validation and testing data.

Additionally, we validate our RF survival prediction in SP1 (RF-SP1) using BraTS18 validation set. We also validate XGBoost survival prediction in SP2 (XGBoost-SP2) using BraTS17 validation dataset. The results are summarized in **Table 1**. Using BraTS17 validation dataset, RF-SP1 model achieves 67.7% accuracy, whereas XGBoost-SP2 model achieves 63.6%. Using BraTS18 validation dataset, RF-SP1 model achieves 46.4%

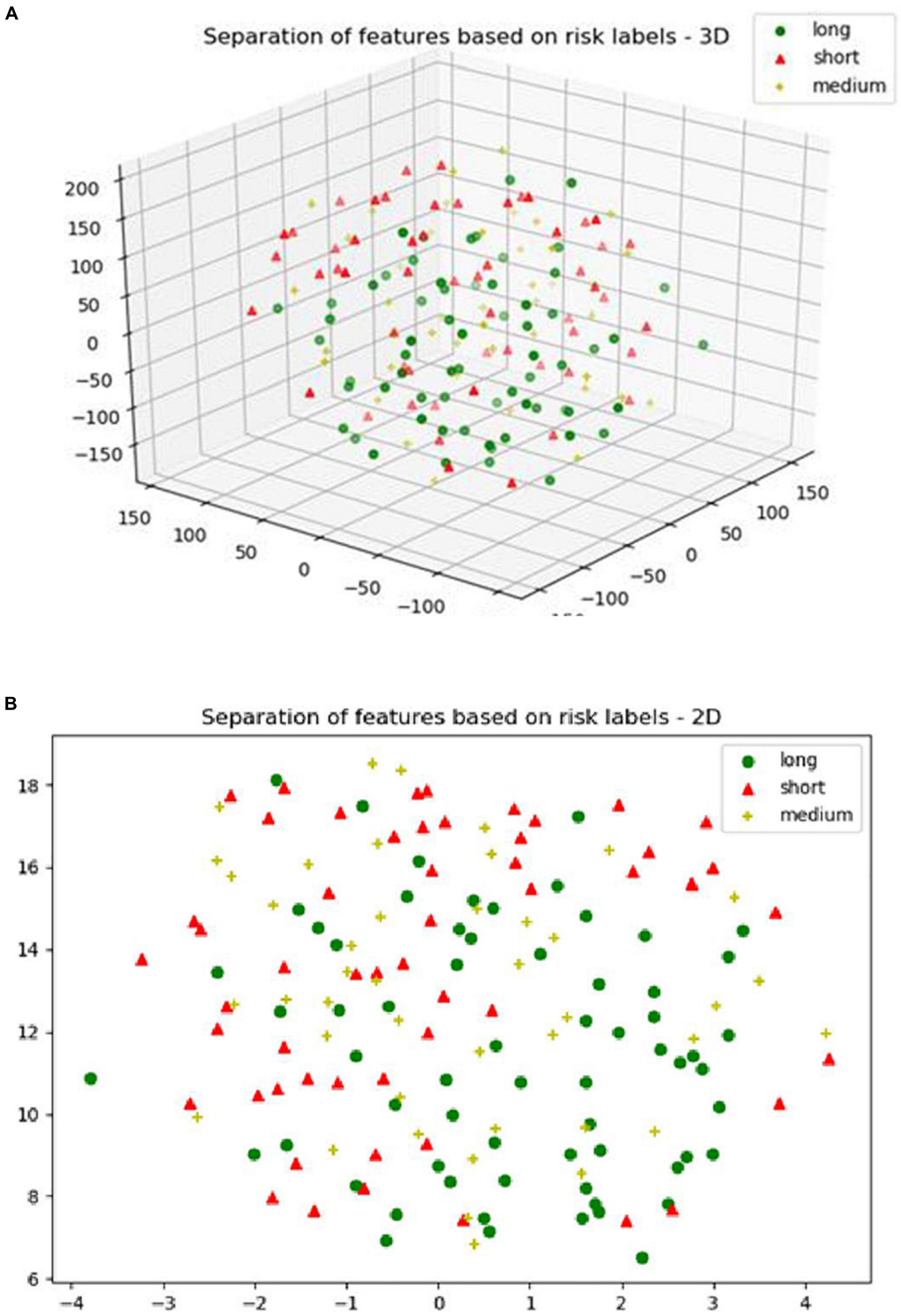


FIGURE 7 | (A) The 3D; and **(B)** the 2D plot of t-distributed stochastic neighbor embedding (t-SNE) of the selected features of SP1 clustered based on the long, medium and short risk labels using BraTS17 training dataset.

accuracy, whereas XGBoost-SP2 model achieves 67.9% accuracy. These results indicate that the XGBoost-SP2 combination performs considerably better than that of RF-SP1 with BraTS18 dataset and reasonably well with BraTS17 dataset, respectively. Note that the ground truth of BraTS17 and BraTS18 validation dataset are not provided. As a result, we have segmented BraTS17

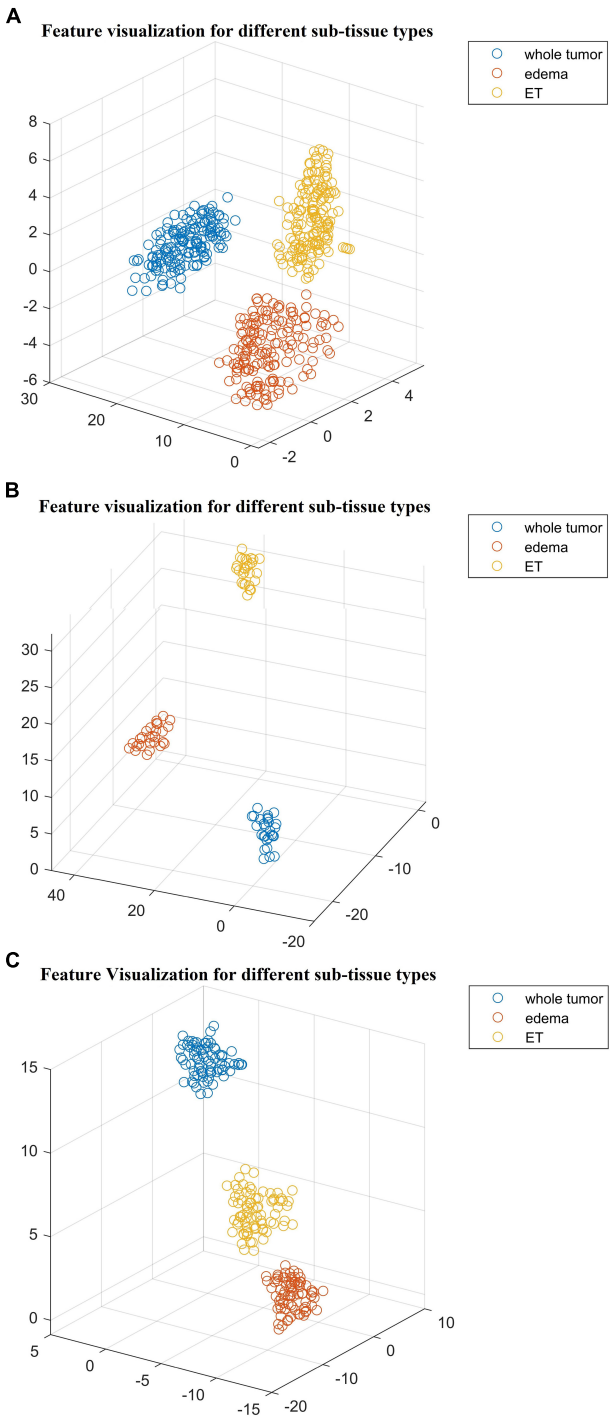


FIGURE 8 | The t-distributed stochastic neighbor embedding (t-SNE) of the selected features of SP2 clustered based on their tissue types using BraTS18 **(A)** training; **(B)** validation; and **(C)** testing. Note that features are clustered based on their origin (subtissue type).

and BraTS18 validation dataset using the semantic label fusion model of CNN and RF (Vidyaratne et al., 2018) and the semantic label fusion of U-Net and FCN, respectively.

Comparison of Survival Prediction With State-of-the-Art Works

Comparison of the proposed survival prediction pipelines SP1 and SP2 with few state-of-the-art methods in literature is discussed next. **Table 5** summarizes the performances of these state-of-the-art models and presents a comparison with our proposed framework (SP2). Chato et al. (2018) propose using histogram features extracted from denoised MR images (by using 2 level Daubechies wavelet transform) in a support vector machine to predict overall survival. Their method achieves a 10-fold cross validation accuracy of 0.667 using BraTS17 training dataset. Kao et al. (2018) extract volumetric, spatial, morphological, and tractographic features from MR images. Feature normalization and selection is performed, and the selected features are trained in a support vector machine model. Their proposed model achieves an accuracy of 0.7 using BraTS18 training dataset and an accuracy of 0.5 using BraTS18 validation dataset. Soltaninejad et al. (2017) utilize volumetric features along with RF to predict overall survival. Their method achieves five-fold cross validation accuracy of 0.638 using BraTS17 training dataset. The results demonstrate that our proposed framework achieves a higher accuracy in overall survival prediction compared to the current-state-of-the-art models applied to the same datasets. Note that, unlike our proposed SP1 and SP2 pipelines, the reported performance for all these other methods in **Table 5** are obtained by the authors themselves. In addition, a comparison between the performance of our segmentation model and state-of-the-art models is illustrated in **Table 6**. Though the abnormal brain tumor tissue segmentation results for other methods in the 2018 Challenge (as shown in **Table 6**) are better than our semantic-label fusion method, our segmentation results are useful to offer the best survival prediction performance in the 2018 BraTS Challenge as shown in **Table 1**.

Modified-SP2

In order to reduce the high dimensionality of the features in SP2 classification and regression steps, we modify SP2 in **Figure 3B** as follows: (1) calculate and rank the feature importance for each classification and regression model; (2) select features that have a relative scaled importance greater than 50%; and (3) train the modified selected features in a new classification and regression training models utilizing XGBoost.

The resulting 30 significant features are applied in the classification step of the modified-SP2. The distribution of these features is as follows: 13 features represent Euler characteristics, 7 features represent volumetric and area-related properties, 4 histogram-graph based features, 5 texture features, and one feature with Age information.

The number of significant features used in the short-, medium-, and long-regression models of the modified-SP2 is 11, 9, and 11, respectively. The distribution of the features in the modified short-regression model are as follows: 2 volumetric and area-related features, 1 histogram-graph

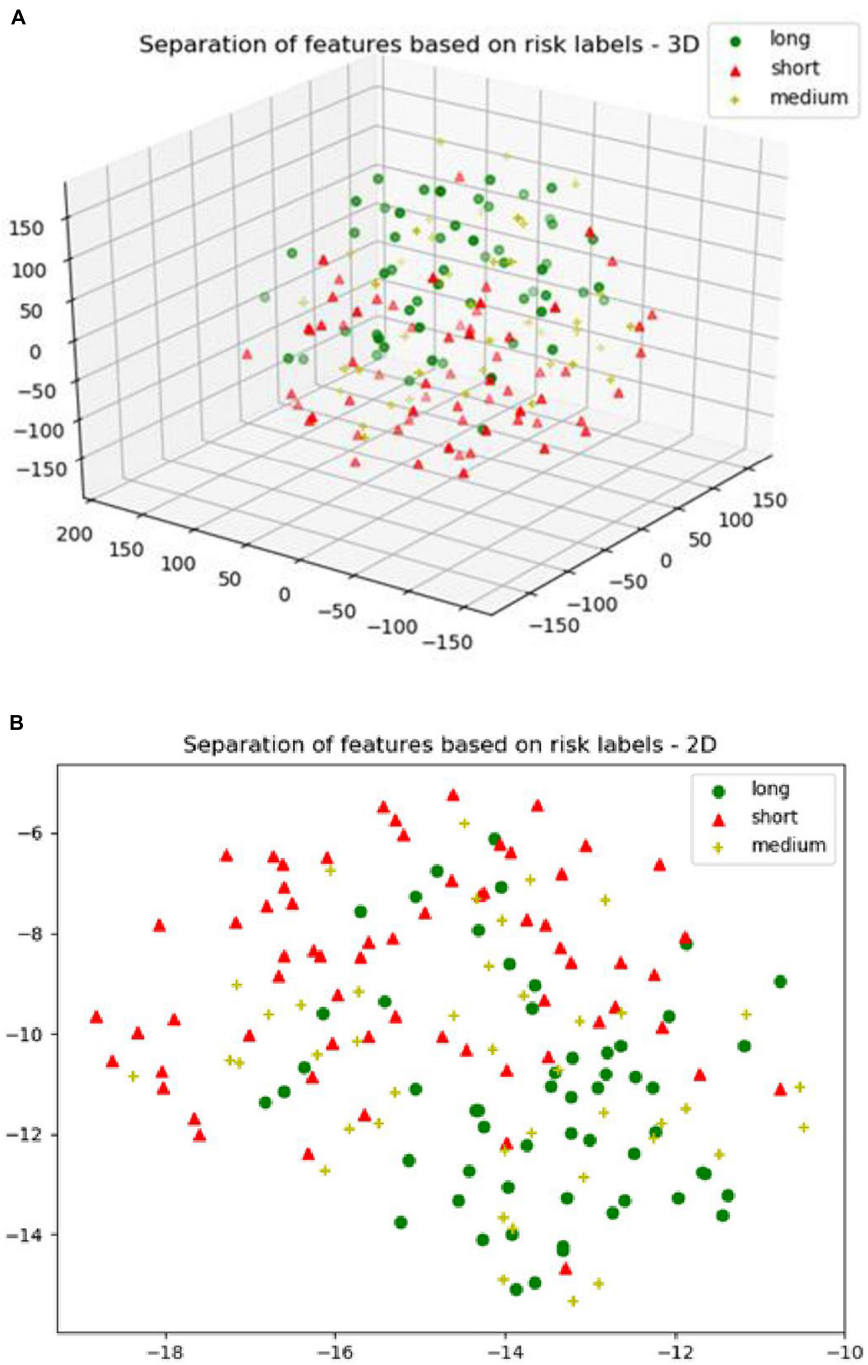


FIGURE 9 | (A) The 3D plot of the t-distributed stochastic neighbor embedding (t-SNE) of the selected features of SP2 clustered based on the long, medium and short risk labels using BraTS18 training dataset. **(B)** The 2D plot of the same training dataset.

based features, 7 texture features, and one feature with Age information. The features employed in the modified med-regression model are 5 volumetric and area-related features, 3 texture features, and Age. Whereas the features of the modified long-regression model are 2 volumetric and area-related features, 8 texture features, and one feature with Age information.

The modified-SP2 achieves cross-validated accuracy of 0.718 as illustrated in **Table 1**. **Table 3** illustrates the statistics of its confusion matrix in the classification training model. **Table 4** illustrates the performance of the modified regression training models. Additionally, the modified-SP2 is validated using BraTS18 validation set and its performance is illustrated in **Table 1**. Note that the different performances of SP2 and

TABLE 5 | Comparison of our proposed survival prediction pipeline with state-of-the-art methods in literature.

References	Algorithm	Validation method	Performance	Dataset
Chato et al., 2018	Histogram features along with SVM	10-fold cross validation	accuracy of 0.667	BraTS17 training dataset
Kao et al., 2018	Volumetric, spatial, morphological, and tractographic features along with SVM	5-fold cross-validation	Accuracy of 0.7	BraTS18 training dataset
Soltaninejad et al., 2017	Volumetric features along with Random Forest	5-fold cross validation	Accuracy of 0.638	BraTS17 training dataset
XGBOOST overall survival prediction model (SP2)	Texture, volumetric, histogram-graph, and Euler features Along with XGBoost	LOOCV	Accuracy of 0.73 and MSE of 91585.51	BraTS18 training dataset
		Validation dataset	Accuracy of 0.679 and MSE of 153466.3	BraTS18 validation dataset

modified-SP2 are almost similar when using the BraTS18 training and validation dataset statistics of each class in SP2 and the modified-SP2 are almost similar. This can be explained by the fact that XGBoost provides L1 and L2 regularization.

Additionally, the modified-SP2 is validated using BraTS18 validation set and its performance is illustrated in **Table 1**.

DISCUSSION AND FUTURE WORKS

This work proposes a novel framework for fully automated deep radiomics-based Glioblastoma segmentation and survival prediction. The overall framework is designed as two-step process where automated tumor segmentation is carried out in the first step, and the segmentation outcome is then used for survival prediction in the second step. The accurate segmentation of abnormal tissue tumor types such as necrosis, edema, and enhancing tissue is critical to ensure robust survival prediction performance. Consequently, several deep learning- and radiomic-feature based segmentation algorithms, and a semantic label fusion are introduced to obtain sufficient segmentation performance. The framework also includes two survival prediction algorithms SP1 and SP2 in step two, represented by the use of feature types, feature selection, regression and classification methods.

The primary survival pipeline (SP1) combines patch-wise CNN based algorithm and radiomics based algorithm using label fusion for segmentation, and applies the RF based survival prediction algorithm to obtain the final output. The second pipeline (SP2) combines U-Net and FCN segmentation with an XGBoost based survival prediction algorithm. As shown in **Figure 1**, the features used in both SP2 and SP1 offers an excellent segmentation of different abnormal tissue type. The functionality of SP2 is further enhanced by using additional features extracted from the subtissues (edema, enhance tumor, and necrosis) and a two-step classification and regression method. Different studies (Pierallini et al., 1998; Lacroix et al., 2001; Maldaun et al., 2004; Jain et al., 2014) correlate between survival prediction in glioblastoma and different subtissues. SP2 shows improvements over our primary survival prediction model (SP1) (Shboul et al., 2017) with LOOCV accuracy increase to 0.73 from 0.67 for training datasets. Whereas the modified-SP2 achieves cross-validation accuracy of 0.718 using the training dataset.

There are a few limitations of the proposed work. First, even though the total number of cases for survival training dataset is 163, both BraTS 2017 and BraTS 2018 required that the data must be divided into three separate survival-group regression models. Consequently, the number of training cases are divided among three models as follows: 65 cases for short-, 42 cases for medium- and 56 cases for long-regression models, respectively. A larger dataset may be required when training each regression model to improve the performance. Second, this study may benefit from additional clinical data such as Gender and Karnofsky Status to strengthen the reliability of the different survival regression and classification models. Finally, the overall survival risk classification performance of the state-of-the-art methods in literature, including the pipelines proposed in this work, may be improved further. The visualization of survival features suggests the difficulty in separating the high dimensional data into the three distinctive risk classes. This suggests the need for further research in novel feature engineering for survival prediction. Following the efficacy of deep radiomics features in the tumor segmentation step, a possible future direction to further improve the risk classification performance may involve use of deep learning methods to learn all possible features in the survival pipeline.

TABLE 6 | Comparison to our proposed with state-of-art models that have used BraTS18 testing dataset.

References	Dice enhanced tumor	Dice whole tumor	Dice tumor core
Semantic-label fusion method (SP2)	0.705	0.857	0.767
Myronenko, 2018	0.766	0.884	0.815
Isensee et al., 2018	0.779	0.878	0.806
Zhou et al., 2018	0.778	0.884	0.796

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2018.html.

ETHICS STATEMENT

The data used in this work is downloaded from publicly available TCIA/TCGA and BRATS websites.

AUTHOR CONTRIBUTIONS

ZS, MA, LV, LP, and KI conceptualized and designed the study, developed the methodology, and analyzed and interpreted the data. ZS, MA, LV, LP, ME, and KI drafted and revised the manuscript. KI acquired the funding.

FUNDING

This work was partially supported by the National Institutes of Health (R01EB020683). This work uses BraTS challenges dataset and has been evaluated using BraTS evaluation platform.

REFERENCES

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* 5:4006. doi: 10.1038/ncomms5006

Ahmed, S., Iftekharuddin, K., Ogg, R., and Laningham, F. (2009). "Efficacy of texture, shape, and intensity features for robust posterior-fossa tumor segmentation in MRI," in *Proceedings of the SPIE 7260, Medical Imaging 2009: Computer-Aided Diagnosis*, San Francisco.

Ayache, A., and Véhel, J. L. (2004). On the identification of the pointwise Hölder exponent of the generalized multifractional brownian motion. *Stoch. Process. Their Appl.* 111, 119–156. doi: 10.1016/j.spa.2003.11.002

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. *Cancer Imaging Arch.* 4, 1–13. doi: 10.1038/sdata.2017.117

Bleeker, F. E., Molenaar, R. J., and Leenstra, S. (2012). Recent advances in the molecular understanding of glioblastoma. *J. Neurooncol.* 108, 11–27. doi: 10.1007/s11060-011-0793-0

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Brown, R., Zlatescu, M., Sijben, A., Roldan, G., Easaw, J., Forsyth, P., et al. (2008). The use of magnetic resonance imaging to noninvasively detect genetic signatures in oligodendroglioma. *Clin. Cancer Res.* 14, 2357–2362. doi: 10.1158/1078-0432.CCR-07-1964

Chato, L., Chow, E., and Latifi, S. (2018). "Wavelet transform to improve accuracy of a prediction model for overall survival time of brain tumor patients based on mri images," in *Proceedings of the IEEE International Conference on Healthcare Informatics*, Piscataway, NJ.

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM).

Cheplygina, V., de Bruijne, M., and Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* 54, 280–296. doi: 10.1016/j.media.2019.03.009

Dong, H., Yang, G., Liu, F., Mo, Y., and Guo, Y. (2017). "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Proceedings of the Annual Conference on Medical Image Understanding and Analysis* (Berlin: Springer).

Emblem, K. E., Nedregaard, B., Nome, T., Due-Tonnessen, P., Hald, J. K., Scheie, D., et al. (2008). Glioma grading by using histogram analysis of blood volume heterogeneity from MR-derived cerebral blood volume maps. *Radiology* 247, 808–817. doi: 10.1148/radiol.2473070571

Guinney, J., Wang, T., Laajala, T. D., Winner, K. K., Bare, J. C., Neto, E. C., et al. (2017). Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. *Lancet Oncol.* 18, 132–142. doi: 10.1016/S1470-2045(16)30560-5

Gutman, D. A., Cooper, L. A., Hwang, S. N., Holder, C. A., Gao, J., Aurora, T. D., et al. (2013). MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* 267, 560–569. doi: 10.1148/radiol.13120118

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Hu, L. S., Ning, S., Eschbacher, J. M., Gaw, N., Dueck, A. C., Smith, K. A., et al. (2015). Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma. *PLoS One* 10:e0141506. doi: 10.1371/journal.pone.0141506

Iftekharuddin, K. M., Jia, W., and Marsh, R. (2003). Fractal analysis of tumor in brain MR images. *Mach. Vis. Appl.* 13, 352–362. doi: 10.1007/s00138-002-0087-9

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *Proceedings of the International MICCAI Brainlesion Workshop* (Berlin: Springer), 234–244.

Islam, A., Iftekharuddin, K. M., Ogg, R. J., Laningham, F. H., and Sivakumar, B. (2008). "Multifractal modeling, segmentation, prediction, and statistical validation of posterior fossa tumors," in *Proceedings of the Medical Imaging 2008: Computer-Aided Diagnosis* (San Francisco: International Society for Optics and Photonics).

Islam, A., Reza, S. M., and Iftekharuddin, K. M. (2013). Multifractal texture estimation for detection and segmentation of brain tumors. *IEEE Trans. Biomed. Eng.* 60, 3204–3215. doi: 10.1109/TBME.2013.2271383

Itakura, H., Achrol, A. S., Mitchell, L. A., Loya, J. J., Liu, T., Westbroek, E. M., et al. (2015). Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. *Sci. Trans. Med.* 7:303ra138. doi: 10.1126/scitranslmed.aaa7582

Jain, R., Poisson, L., Narang, J., Gutman, D., Scarpace, L., Hwang, S. N., et al. (2013). Genomic mapping and survival prediction in glioblastoma: molecular subclassification strengthened by hemodynamic imaging biomarkers. *Radiology* 267, 212–220. doi: 10.1148/radiol.12120846

Jain, R., Poisson, L. M., Gutman, D., Scarpace, L., Hwang, S. N., Holder, C. A., et al. (2014). Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology* 272, 484–493. doi: 10.1148/radiol.14131691

Johnson, D. R., Leeper, H. E., and Uhm, J. H. (2013). Glioblastoma survival in the United States improved after food and drug administration approval of bevacizumab: a population-based analysis. *Cancer* 119, 3489–3495. doi: 10.1002/cncr.28259

Kao, P.-Y., Ngo, T., Zhang, A., Chen, J., and Manjunath, B. (2018). Brain tumor segmentation and tractographic feature extraction from structural MR images for overall survival prediction. *arXiv*

Kirienko, M., Lozza, L., Cozzi, L., Gennaro, N., Rossi, A., Voulaz, E., et al. (2018). EP-1362: random forest analysis to predict disease-free survival using FDG-PET and CT in lung cancer. *Radiother. Oncol.* 127, S743–S744.

Lacroix, M., Abi-Said, D., Fourney, D. R., Gokaslan, Z. L., Shi, W., DeMonte, F., et al. (2001). A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J. Neurosurg.* 95, 190–198. doi: 10.3171/jns.2001.95.2.0190

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J., et al. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci. Rep.* 7:10353. doi: 10.1038/s41598-017-10649-8

Leung, T., and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vis.* 43, 29–44.

Liu, Y., Xu, X., Yin, L., Zhang, X., Li, L., and Lu, H. (2017). Relationship between glioblastoma heterogeneity and survival time: an MR imaging texture analysis. *Am. J. Neuroradiol.* 38, 1695–1701. doi: 10.3174/ajnr.A5279

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Piscataway, NJ.

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., et al. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* 114, 97–109.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Macyszyn, L., Akbari, H., Pisapia, J. M., Da, X., Attiah, M., Pigrish, V., et al. (2015). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro Oncol.* 18, 417–425. doi: 10.1093/neuonc/nov127

Maldaun, M. V., Suki, D., Lang, F. F., Prabhu, S., Shi, W., Fuller, G. N., et al. (2004). Cystic glioblastoma multiforme: survival outcomes in 22 cases. *J. Neurosurg.* 100, 61–67. doi: 10.3171/jns.2004.100.1.0061

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 3D Vision (3DV), 2016 Fourth International Conference*, (Piscataway, NJ: IEEE).

Mlynarski, P., Delingette, H., Criminisi, A., and Ayache, N. (2018). Deep learning with mixed supervision for brain tumor segmentation. *arXiv*

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *Proceedings of the International MICCAI Brainlesion Workshop* (Berlin: Springer), 311–320. doi: 10.1007/978-3-030-11726-9_28

Ohgaki, H. (2005). Genetic pathways to glioblastomas. *Neuropathology* 25, 1–7. doi: 10.1111/j.1440-1789.2004.00600.x

Passamonti, F., Giorgino, T., Mora, B., Guglielmelli, P., Rumi, E., Maffioli, M., et al. (2017). A clinical-molecular prognostic model to predict survival in patients with post polycythemia vera and post essential thrombocythemia myelofibrosis. *Leukemia* 31:2726. doi: 10.1038/leu.2017.169

Pierallini, A., Bonamini, M., Pantano, P., Palmeggiani, F., Raguso, M., Osti, M., et al. (1998). Radiological assessment of necrosis in glioblastoma: variability and prognostic value. *Neuroradiology* 40, 150–153. doi: 10.1007/s002340050556

Pope, W. B., Sayre, J., Perlina, A., Villablanca, J. P., Mischel, P. S., and Cloughesy, T. F. (2005). MR imaging correlates of survival in patients with high-grade gliomas. *Am. J. Neuroradiol.* 26, 2466–2474.

Rathore, S., Akbari, H., Rozycki, M., Bakas, S., and Davatzikos, C. (2016). *Nimg-20. Imaging Pattern Analysis Reveals Three Distinct Phenotypic Subtypes of GBM With Different Survival Rates*. Oxford: Oxford University Press.

Reza, S., and Iftekharuddin, K. (2014). "Multi-fractal texture features for brain tumor and edema segmentation," in *Proceedings of the Medical Imaging 2014 Computer-Aided Diagnosis* (Bellingham, WA: International Society for Optics and Photonics).

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, eds N. Navab, J. Hornegger, W. Wells, and A. Frangi, (Cham: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Shah, M. P., Merchant, S., and Awate, S. P. (2018). "MS-Net: mixed-supervision fully-convolutional networks for full-resolution segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer), 379–387. doi: 10.1007/978-3-030-00937-3_44

Shboul, Z. A., Vidyaratne, L., Alam, M., and Iftekharuddin, K. M. (2017). "Glioblastoma and survival prediction," in *Proceedings of the International MICCAI Brainlesion Workshop* (Berlin: Springer), 358–368.

Shouval, R., Ruggeri, A., Labopin, M., Mohty, M., Sanz, G., Michel, G., et al. (2017). An Integrative scoring system for survival prediction following umbilical cord blood transplantation in acute leukemia. *Clin. Cancer Res.* 23, 6478–6486. doi: 10.1158/1078-0432.CCR-17-0489

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*

Soltaninejad, M., Zhang, L., Lambrou, T., Yang, G., Allinson, N., and Ye, X. (2017). "MRI brain tumor segmentation and patient survival prediction using random forests and fully convolutional networks," in *Proceedings of the International MICCAI Brainlesion Workshop* (Berlin: Springer), 204–215. doi: 10.1007/978-3-319-75238-9_18

Turner, K., Mukherjee, S., and Boyer, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Inform. Inference* 3, 310–344. doi: 10.1093/imaiai/iau011

Vallières, M., Freeman, C. R., Skamene, S. R., and El Naqa, I. (2015). A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* 60:5471. doi: 10.1088/0031-9155/60/14/5471

Vartanian, A., Singh, S. K., Agnihotri, S., Jalali, S., Burrell, K., Aldape, K. D., et al. (2014). GBM's multifaceted landscape: highlighting regional and microenvironmental heterogeneity. *Neuro Oncol.* 16, 1167–1175. doi: 10.1093/neuonc/nou035

Vidyaratne, L., Alam, M., Shboul, Z., and Iftekharuddin, K. (2018). "Deep learning and texture-based semantic label fusion for brain tumor segmentation," in *Proceedings of the Medical Imaging 2018: Computer-Aided Diagnosis* (San Francisco: International Society for Optics and Photonics).

Yang, D., Korogi, Y., Sugahara, T., Kitajima, M., Shigematsu, Y., Liang, L., et al. (2002). Cerebral gliomas: prospective comparison of multivoxel 2D chemical-shift imaging proton MR spectroscopy, echoplanar perfusion and diffusion-weighted MRI. *Neuroradiology* 44, 656–666. doi: 10.1007/s00234-002-0816-9

Yang, D., Rao, G., Martinez, J., Veeraraghavan, A., and Rao, A. (2015). Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. *Med. Phys.* 42, 6725–6735. doi: 10.1118/1.4934373

Zhao, X., Wu, Y., Song, G., Li, Z., Fan, Y., and Zhang, Y. (2016). "Brain tumor segmentation using a fully convolutional neural network with conditional random fields," in *Proceedings of the International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Berlin: Springer), 75–87. doi: 10.1007/978-3-319-55524-9_8

Zhou, C., Chen, S., Ding, C., and Tao, D. (2018). "Learning contextual and attentive information for brain tumor segmentation," in *Proceedings of the International MICCAI Brainlesion Workshop* (Berlin: Springer), 497–507. doi: 10.1007/978-3-030-11726-9_44

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 05 November 2019
doi: 10.3389/fnins.2019.01182

Divide and Conquer: Stratifying Training Data by Tumor Grade Improves Deep Learning-Based Brain Tumor Segmentation

Michael Rebsamen[1,2]*, *Urspeter Knecht*[3], *Mauricio Reyes*[4], *Roland Wiest*[1], *Raphael Meier*[1†] and *Richard McKinley*[1†]

[1] Support Center for Advanced Neuroimaging (SCAN), University Institute of Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland, [2] Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland, [3] Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland, [4] Healthcare Imaging A.I. Lab, Insel Data Science Center, Inselspital, Bern University Hospital, Bern, Switzerland

OPEN ACCESS

Edited by:
Bjoern Menze,
Technical University of Munich, Germany

Reviewed by:
Roberto Viviani,
University of Innsbruck, Austria
Benedikt Wiestler,
Technical University of Munich, Germany

***Correspondence:**
Michael Rebsamen
michael.rebsamen@insel.ch

[†]These authors share senior authorship

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 30 April 2019
Accepted: 18 October 2019
Published: 05 November 2019

Citation:
Rebsamen M, Knecht U, Reyes M, Wiest R, Meier R and McKinley R (2019) Divide and Conquer: Stratifying Training Data by Tumor Grade Improves Deep Learning-Based Brain Tumor Segmentation. Front. Neurosci. 13:1182. doi: 10.3389/fnins.2019.01182

It is a general assumption in deep learning that more training data leads to better performance, and that models will learn to generalize well across heterogeneous input data as long as that variety is represented in the training set. Segmentation of brain tumors is a well-investigated topic in medical image computing, owing primarily to the availability of a large publicly-available dataset arising from the long-running yearly Multimodal Brain Tumor Segmentation (BraTS) challenge. Research efforts and publications addressing this dataset focus predominantly on technical improvements of model architectures and less on properties of the underlying data. Using the dataset and the method ranked third in the BraTS 2018 challenge, we performed experiments to examine the impact of tumor type on segmentation performance. We propose to stratify the training dataset into high-grade glioma (HGG) and low-grade glioma (LGG) subjects and train two separate models. Although we observed only minor gains in overall mean dice scores by this stratification, examining case-wise rankings of individual subjects revealed statistically significant improvements. Compared to a baseline model trained on both HGG and LGG cases, two separately trained models led to better performance in 64.9% of cases ($p < 0.0001$) for the tumor core. An analysis of subjects which did not profit from stratified training revealed that cases were missegmented which had poor image quality, or which presented clinically particularly challenging cases (e.g., underrepresented subtypes such as IDH1-mutant tumors), underlining the importance of such latent variables in the context of tumor segmentation. In summary, we found that segmentation models trained on the BraTS 2018 dataset, stratified according to tumor type, lead to a significant increase in segmentation performance. Furthermore, we demonstrated that this gain in segmentation performance is evident in the case-wise ranking of individual subjects but not in summary statistics. We conclude that it may

be useful to consider the segmentation of brain tumors of different types or grades as separate tasks, rather than developing one tool to segment them all. Consequently, making this information available for the test data should be considered, potentially leading to a more clinically relevant BraTS competition.

Keywords: magnetic resonance imaging, brain tumors, automatic segmentation, deep learning, training strategy, data stratification

1. INTRODUCTION

Gliomas are primary brain tumors which arise from glial cells. According to the World Health Organization (WHO) classification of tumors of the central nervous system (CNS) (Louis et al., 2016), they can be grouped into different tumor grades based on the underlying histology and molecular characteristics. Increasing tumor grade indicates the increasing malignancy of the tumor. Glioma are managed depending on grade, with treatment strategies ranging from tumor resection followed by combined radio- and chemotherapy to a "watch and wait" approach (Stupp et al., 2005; Grier, 2006). Glioblastoma are the most aggressive type of glioma (WHO grade IV) and make up 45% of all gliomas (Ostrom et al., 2014). The prime imaging technique in brain tumor diagnostics is Magnetic Resonance Imaging (MRI) (Essig et al., 2012). Standard acquisition protocols used to perform initial diagnosis and treatment monitoring include T1-weighted, T1-weighted gadolinium-enhanced, T2-weighted, and T2-weighted with fluid attenuated inversion recovery (FLAIR) sequences (Wen et al., 2010; Ellingson et al., 2015). The typical radiological appearance of a glioblastoma features a disrupted blood-brain barrier causing ring-enhancing lesions with central necrosis and peritumoral edema. In contrast, low-grade astrocytic tumors exhibit typically no contrast enhancement and are missing central necrosis (Pierallini et al., 1997).

In the case of glioblastoma, recent studies led to the discovery of a profound genetic heterogeneity among, and even within, tumors (Verhaak et al., 2010; Sottoriva et al., 2013). It has been shown that the underlying genetic and molecular heterogeneity can be associated with variations in imaging phenotype such as changes in tumor compartment volumes (Lai et al., 2013; Grossmann et al., 2016), contrast enhancement (Carrillo et al., 2012; Treiber et al., 2018), radiomic signatures (Gevaert et al., 2014), and tumor location (Carrillo et al., 2012; Ellingson et al., 2012). The imaging appearance of glioblastoma can further be altered by treatment causing radiation necrosis (Mullins et al., 2005) and pseudoprogression and -response (Hygino da Cruz et al., 2011), respectively. As a consequence, a machine learning segmentation algorithm needs to be capable of generalizing across this heterogeneity of glioblastoma imaging phenotypes.

Brain tumor segmentation is a well-investigated topic with a vast amount of available methods and yearly organized MICCAI Brain Tumor Segmentation (BraTS) Challenges since the year 2012 (Menze et al., 2015; Bakas et al., 2017c), serving as a public platform for algorithm comparison. With the rise of deep learning, brain tumor segmentation methods experienced significant gains in performance (Bakas et al., 2018). One of the central promises of deep learning methods is that they can be fed with raw data and are capable of automatically uncovering the underlying representation relevant for the task at hand (e.g., segmentation) from that data (LeCun et al., 2015). As a consequence, the time-consuming and error-prone manual engineering of features traditionally used in machine learning has been rendered obsolete. Recently, it was shown for vision tasks that model performance increases logarithmically based on volume of training data (Sun et al., 2017). This aligns with the general notion that more training data leads to a better generalization of a machine learning algorithm. Within the context of BraTS Challenges, deep learning methods are usually trained *ad hoc* on all of the available data, disregarding underlying latent factors such as genetic characteristics or even tumor grades. Although the tumor type is available to the challenge participants for the training data, this information is withheld for the validation and test data. Since part of the BraTS dataset is coming from *The Cancer Imaging Archive* (TCIA) (Bakas et al., 2017a,b,c), additional relevant information such as e.g., patient's gender, mutation subtypes [Isocitrate dehydrogenase (IDH), 1p19q co-deletion] and methylation status of MGMT-promotor could potentially be added as well.

The metric of choice for algorithm comparison in biomedical image segmentation challenges is the Dice coefficient, which was used in 92% of the 383 segmentation tasks reported in Maier-Hein et al. (2018). Predominantly, the Dice coefficient is reported in terms of summary statistics (mean/median) over patient cases and model comparison is performed on the basis of such summary statistics (metric-based ranking). Recently, the BraTS Challenge adopted a case-based ranking scheme. While metric-based rankings lead to more robust rankings than case-based rankings (Maier-Hein et al., 2018), it can be argued that distinct performance differences for individual patients may be obfuscated.

We hypothesize that deep learning methods for brain tumor segmentation can be significantly improved by taking into account latent factors along with tumor image appearance during model training. The purpose of this study is to demonstrate the impact of including prior knowledge of a particular latent factor (tumor grade) on the performance of a recently published, top-ranked deep learning method (McKinley et al., 2019a). Furthermore, the impact is studied employing both a metric-based and case-based rank analysis.

The idea of leveraging prior information about tumor grades to improve segmentation has been presented as an extended abstract to the *International Conference on Medical Imaging with Deep Learning (MIDL)* along with preliminary results (Meier et al., 2019).

2. MATERIALS AND METHODS

2.1. Study Data

The study is based on publicly-available data of the BraTS 2018 Challenge (Menze et al., 2015; Bakas et al., 2017c). In particular, the training dataset was used, which includes 75 patients with low-grade glioma (LGG) and 210 patients with high-grade glioma (HGG). The imaging data encompasses four MR image sequences (T1-weighted, T1-weighted with contrast agent, T2-weighted, and T2-weighted FLAIR sequences), which are part of the consensus recommendations for a standardized brain tumor imaging protocol in clinical trials (Ellingson et al., 2015). The imaging data stem from 19 different institutions, which relied on different MR scanners and acquisition protocols. Manual segmentations of three tumor compartments were available: contrast-enhancing tumor, non-enhancing/necrosis combined, and edema. The regions which were considered for evaluation in the BraTS 2018 challenge as well as in the study at hand were: contrast-enhancing tumor, tumor core (all compartments except edema), and whole tumor (all compartments). More details on the preprocessing and the evolution of the BraTS dataset can be found in Bakas et al. (2017c).

2.2. Automatic Segmentation

The network architecture used for the automatic segmentation is equivalent to the model ranked third in the BraTS 2018 challenge (McKinley et al., 2019a). In brief, it is a U-net-style structure with densely connected blocks of dilated convolutions. The segmentation is performed slice-wise where the input data includes the two neighboring slices from below and above from all four image modalities (i.e., input dimension is $batch \times 4 \times 5 \times 192 \times 192$). The final segmentation is the result of ensembling the predictions from all three directions (sagittal, axial, and coronal).

In a pre-processing step, the data are first normalized to zero mean and unit variance. Data augmentation consists of a combination of randomly flipping the images along the midline and random rotations [$angle \sim U(-15, +15)$] around all principal axis. Additionally, the standardized voxel intensities are randomly shifted [$amount \sim N(0, 0.5)$] and scaled [$factor \sim N(1, 0.2)$].

The networks were trained with a *focal loss* function, *RMSprop* as optimizer with a cosine-annealing learning rate schedule, and a batch-size of two.

2.3. Stratified Model Training

Three different models were trained independently, each with a five-fold cross-validation: A baseline model with all available training data (number of samples $N = 285$), an HGG-only model ($N = 210$), and an LGG-only model ($N = 75$). Network architecture and hyperparameters were the same for all models which were trained on a *Nvidia GeForce GTX 1080 Ti* GPU with 11GB memory over 80 epochs. Qualitatively, the performance on the validation-set was saturating with no observable overfitting (see **Figure S1**).

2.4. Statistical Analysis

The statistical analysis was performed using *R* with the *stats* package version 3.5.1 (R Core Team, 2018). For comparison of spatial overlap of estimated tumor segmentations with manual ground truth data, the Dice coefficient was used. Segmentation performance in terms of Dice coefficient of the different deep learning models was summarized by descriptive statistics (median, interquartile range). Case-based rank analysis included computation of percentage of improved patient cases for given pairing of deep learning models. The stratified models were compared to the baseline by means of paired difference tests: differences between the cross-validated classifiers were examined on HGG cases only, on LGG cases only, and on the whole dataset (using the combined results of the stratified LGG and HGG classifiers). Non-parametric tests were employed due to the rank-based form of the data. The significance level of the analysis was set to α=0.05.

3. RESULTS

3.1. Quantitative Analysis

Summary statistics for the segmentation performance in terms of Dice coefficient are shown in **Figure 1**. The baseline model reached a median Dice of 0.841 ($1.5 \times$ IQR $=$ 0.465–1.000) for the contrast enhancing compartment, 0.899 (0.554–1.000) for the core, and 0.920 (0.786–1.000) for the whole tumor. Comparable, the combined results from the separately trained HGG/LGG models were 0.838 (0.415–1.000) for contrast enhancing, 0.902 (0.584–1.000) for core, and 0.916 (0.800–1.000) for tumor.

The combined results of the two separately trained models showed an improvement for the segmentation of the tumor core in 64.9% ($p < 0.0001$) of the subjects compared to the baseline model (**Table 1**). No statistically significant changes were observed for the other compartments. This performance gain originates primarily from the HGG cases where 70.3% of the subjects showed an improved segmentation for the tumor core and 58.5% of the subjects also for the contrast enhancing compartment. From the 183 subjects that showed an improved segmentation of the core, 26 increased by a Dice of 0.1 or more. Conversely, from the 99 subjects with a declined performance, 21 decreased by a Dice of -0.1 or more (**Figure 2**).

3.2. Qualitative Analysis of Selected Cases

From **Table 1** it is evident that, especially for high-grade glioma, stratified training leads to improved segmentation performance. In order to further investigate this aspect, a visual review of selected cases was performed. To identify cases mostly affected by the stratified training, Dice coefficients between the segmentations of the two models (baseline vs. HGG) were calculated. Cases with a Dice agreement < 0.8 of the tumor core between the baseline and stratified models were selected for a qualitative manual inspection followed by a review with a board-certified neuroradiologist with more than 8 years of experience in brain tumor diagnostics. In order to render the visual review more systematic, we define three categories of causes for variability in tumor segmentation performance: 1. The input data generated by the imaging process, which is affected by the idiosyncrasies of the MR scanner, potential image artifacts and patient motion, and image preprocessing. 2. The manual ground truth segmentation. 3. The tumor

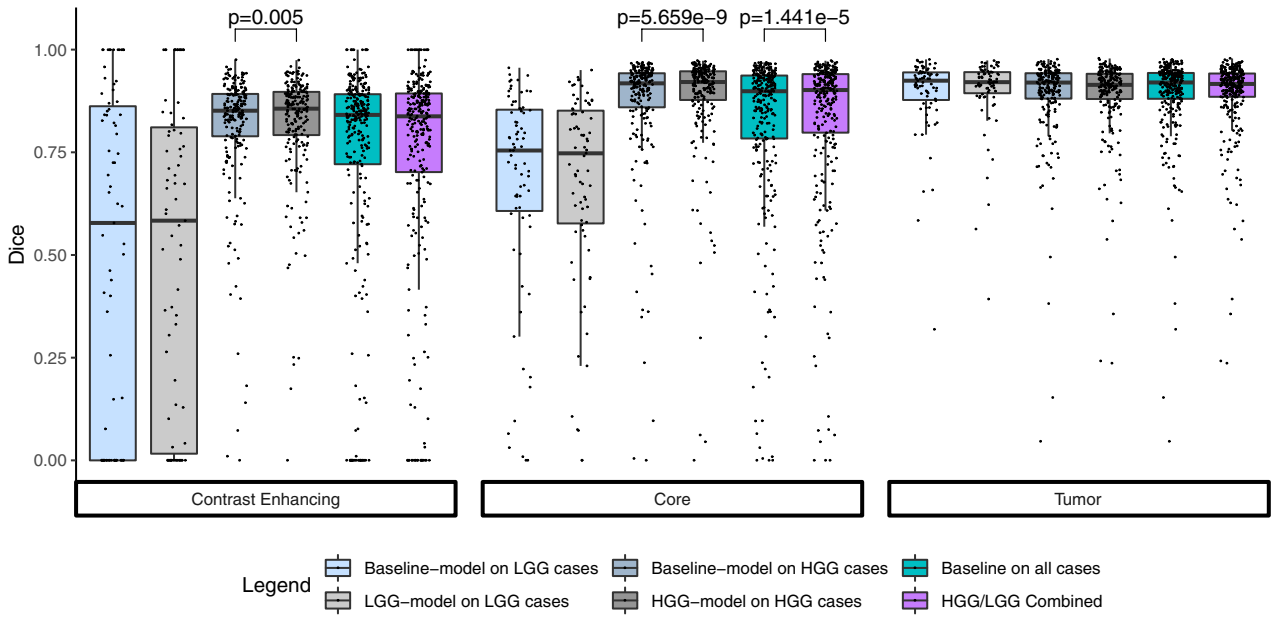


FIGURE 1 | Summary statistics for the segmentation of the three compartments by means of a Tukey boxplot. *p*-values indicate statistically significant ($p < 0.05$) improvements determined by a one-sided Wilcoxon signed rank test.

phenotype (e.g., IDH-mutant tumor, presence of intratumoral hemorrhage, or cystic components) which causes distinctively different image appearances.

In **Figure 3** the obtained Dice coefficients between the segmentation results of the HGG model for the tumor core and the ground truth were plotted against the Dice coefficients between the results of the HGG model and the segmentation of the baseline model, which was trained on all available data. We can broadly define four different territories in the scatterplot: The upper right corner which contains cases for which both models achieved high segmentation performance. If we move to the upper left corner, we encounter cases for which the HGG model achieved high segmentation performance with discrepancies when compared to the results of the baseline model. If we move from the upper right corner to the lower right corner, we encounter cases for which the HGG model agreed with the segmentation result of the baseline model but did not agree with the ground truth result. Finally, the lower left corner contains cases for which the segmentation results of the HGG model did neither agree with the ground truth nor with the segmentation of the baseline model. The corresponding scatterplots for the other two compartments can be found in **Figures S2, S3**. The identified outlier cases are listed in **Table 2** with the segmentation performance of the two models and an assessment category. Below we present the observations based on a visual review for a selection of the identified outliers. Visualizations for the remaining outliers can be found in **Figures S4–S14**.

Brats18_2013_21_1 (**Figure 4**). The baseline model provided superior performance for segmenting the tumor core in this HGG example. The lesion exhibits a large non-enhancing tumor mass (typically seen in LGG) and we speculate that the presence of LGG cases in the baseline model led to the improved tumor core segmentation performance when compared to the HGG model's result. The appearance of the tumor is further complicated by the presence of cystic components, which exhibit a homogeneous signal that is strongly hypointense in T1-weighted and hyperintense in T2-weighted images.

TABLE 1 | Ratio in % of better performing subjects compared to baseline.

	CE		Core		Tumor	
	% Subjects	*p*	% Subjects	*p*	% Subjects	*p*
LGG vs. Baseline	41.7	0.877	49.3	0.454	54.7	0.208
HGG vs. Baseline	**58.4**	**0.005**	**70.3**	**5.659e-09**	46.7	0.877
HGG/LGG vs. Baseline	54.6	0.127	**64.9**	**1.441e-05**	48.8	0.725

Statistical significance is determined by a one-sided Wilcoxon signed rank test. Bold numbers indicate statistically significant ($p < 0.05$) results. CE: contrast enhancing.

Brats18_2013_25_1 (**Figure 5**). Both models failed to segment the tumor core for this HGG case. The tumor core contains strongly hypointense areas in the T2-weighted and FLAIR images with corresponding heterogeneous signal intensity in the T1-weighted image. When considering the T1/T1c-weighted images, one can observe the presence of recruited blood vessels. This image appearance may indicate the presence of an intratumoral hemorrhage.

Brats18_CBICA_AXJ_1 (**Figure 6**). The segmentation of the core from the HGG model is closer to ground truth. The tumor was indicated to be an HGG. However, the provided ground truth segmentation seems to be missing part of the

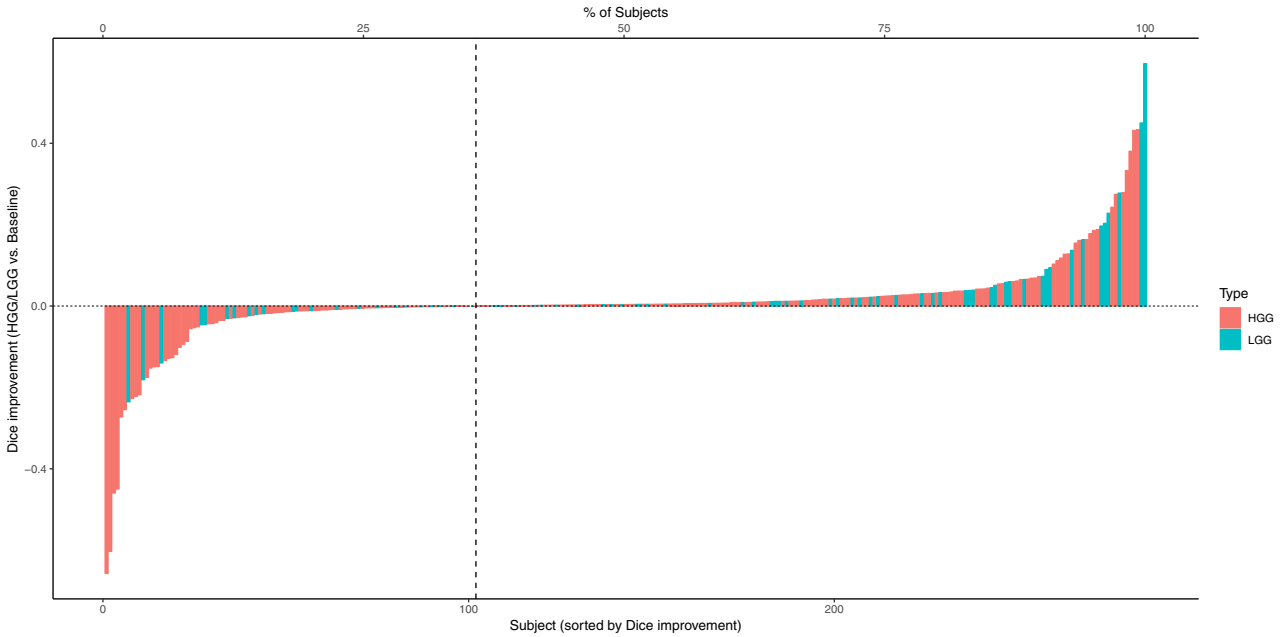


FIGURE 2 | Absolute change of Dice coefficients of the tumor core for each subject. Positive changes were observed for the subjects on the right side of the dashed vertical line.

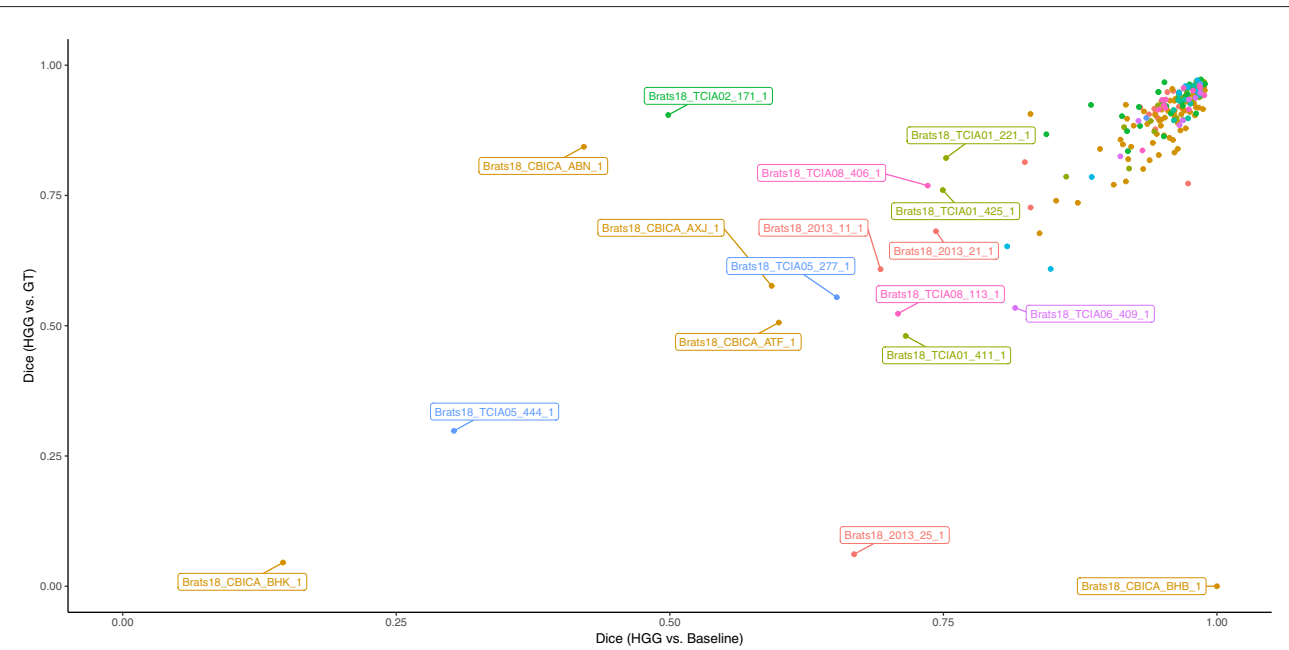


FIGURE 3 | Performance of the HGG-only model for the tumor core (y-axis) and agreement with the baseline model (x-axis). Subjects with a label were visually reviewed. Colors indicate the center (2013, CBICA, TCIA01-08).

tumor mass in the frontal lobe. Furthermore, we argue that a large part of the lesion corresponds to non-enhancing tumor rather than edema. We base this assumption on the heterogeneous appearance in the T2-weighted images and more importantly the strong cortical space-occupying effect together with a distortion of the gray/white matter junction. In contrast, edema would preserve the gray/white matter junction as well as the cortical ribbon and propagate along the white matter fiber tracts. A possible alternative for ground truth is shown in **Figure 6**.

TABLE 2 | Performance of selected cases for the two models.

Subject	Assessment	Dice Baseline-model			Dice HGG-model		
		CE	Core	Tumor	CE	Core	Tumor
Brats18_2013_11_1	1	0.14	0.45	0.90	0.17	0.61	0.89
Brats18_2013_21_1	3	0.80	0.83	0.94	0.76	0.68	0.94
Brats18_2013_25_1	3	0.18	0.10	0.90	0.25	0.06	0.90
Brats18_CBICA_ABN_1	2	0.84	0.41	0.82	0.79	0.84	0.83
Brats18_CBICA_ATF_1	3	0.69	0.73	0.69	0.65	0.51	0.63
Brats18_CBICA_AXJ_1	2	0.79	0.35	0.90	0.79	0.58	0.90
Brats18_CBICA_BHB_1	2	0.00	0.00	0.15	0.00	0.00	0.24
Brats18_CBICA_BHK_1	2	0.01	0.00	0.05	0.25	0.05	0.24
Brats18_TCIA01_221_1	2	0.76	0.88	0.95	0.48	0.82	0.95
Brats18_TCIA01_411_1	1	0.07	0.24	0.71	0.23	0.48	0.64
Brats18_TCIA01_425_1	–	0.26	0.58	0.75	0.68	0.76	0.78
Brats18_TCIA02_171_1	2	0.89	0.47	0.95	0.89	0.90	0.95
Brats18_TCIA04_343_1	2	0.69	0.73	0.74	0.59	0.61	0.66
Brats18_TCIA05_277_1	3	0.42	0.37	0.85	0.56	0.55	0.90
Brats18_TCIA05_444_1	3	0.39	0.96	0.94	0.54	0.30	0.89
Brats18_TCIA06_409_1	–	0.52	0.53	0.89	0.50	0.53	0.86
Brats18_TCIA08_113_1	1	0.91	0.36	0.97	0.79	0.52	0.92
Brats18_TCIA08_406_1	1	0.65	0.63	0.88	0.68	0.77	0.90

Assessment after a qualitative review with a neuroradiologist. Assessment 1: Issue with input image quality, 2: Possible problem with ground truth, 3: Special phenotype, GT: ground truth, CE: contrast enhancing.

Brats18_CBICA_BHB_1 (**Figure 7**). Both models failed completely to segment the lesion for this HGG case. However, the provided ground truth segmentation seems to overestimate the presence of edema. While we agree on the whole tumor segmentation, we argue that the present T2-weighted hyperintensity indicates the presence of non-enhancing tumor rather than edema. Similarly to case Brats18_CBICA_AXJ_1 the gray/white matter junction is distorted. This is especially evident when considering the unaffected contralateral hemisphere. The poor segmentation performance of both models might be the result of an under-representation of training samples with such a subtle tumor core which is potentially ambiguously labeled in other cases as well.

Brats18_TCIA01_221_1 (**Figure 8**). The baseline model provided the better tumor core segmentation for this HGG case. However, when comparing the segmentation of the contrast-enhancing tumor of the HGG model, we argue that the ground truth segmentation slightly undersegments it. This is clearly visible for the enhancing rim next to the midline.

Brats18_TCIA01_425_1 (**Figure 9**). The baseline model underestimated the subtle contrast-enhancement of this HGG case. We can speculate that in the situation of subtle enhancements the baseline model was biased more toward segmenting a tumor core with small enhancing foci, whereas the HGG model was capable of delineating the full extent of the contrast-enhancement.

Brats18_TCIA05_444_1 (**Figure 10**). The baseline model provided a better segmentation than the HGG model for this case. The tumor was indicated to be an HGG. The location of the tumor in the frontal lobe and its appearance exhibiting focal contrast enhancements and a large non-enhancing tumor mass are suspicious of a potential IDH-mutant glioblastoma. This would imply that it initially emerged from an LGG (called secondary glioblastoma). Applying the LGG model to the case significantly outperforms the HGG model (**Figure 10**), which would support the hypothesis of a mutated LGG.

While the previous analysis of cases was to some extent speculation, we can nevertheless condense three main, factual observations from it: First, individual segmentation results are strongly affected by the composition of the segmentation model's training data. Second, depending on the underlying factors that caused a given image appearance and segmentation ground truth, a given subset of the training data can actually improve the segmentation result compared to a baseline trained on all data. Third, disagreement (or joint failure) among segmentation models trained on different subsets of training data (**Figure 3**) may actually help in the identification of these underlying factors. Among the manually reviewed 18 cases with a large deviation between the two models, we observed issues with the input images (4 cases), potentially arguable ground truth (7 cases), and special imaging phenotypes (5 cases). Arguable ground truth is often attributed to edema that could be labeled as tumor core instead. Edema typically propagates along white matter and spares cortical ribbons as well as deep gray matter structures (Pope et al., 2005), while non-enhancing tumor leads to a distortion of the gray/white matter junction [cf. BRATS18_CBICA_BHB_1 (**Figure 7**) FLAIR with the case presented in Figure 3 of Lasocki and Gaillard, 2019].

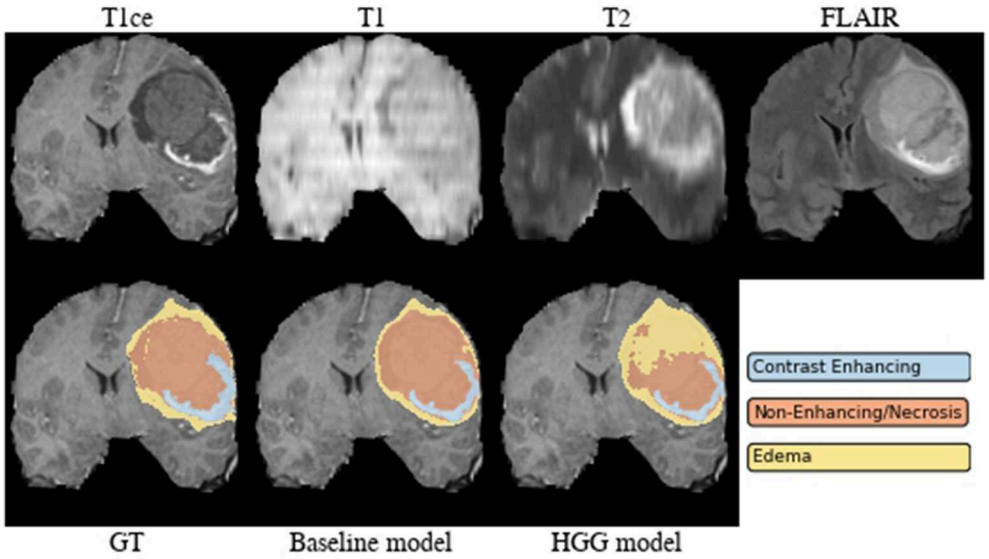


FIGURE 4 | Brats18_2013_21_1.

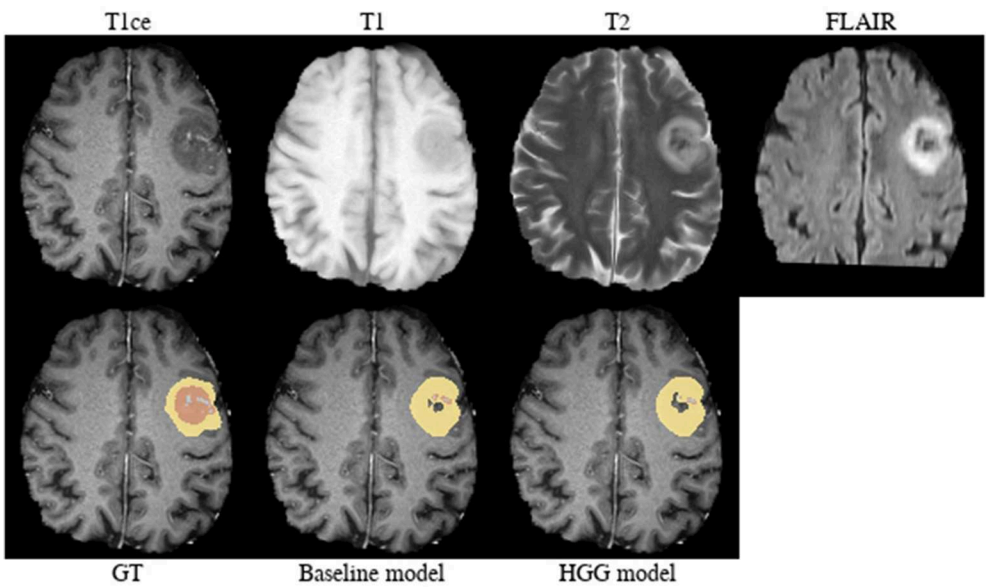


FIGURE 5 | Brats18_2013_25_1.

4. DISCUSSION

The title of the manuscript contains the phrase "Divide and Conquer," where "Divide" refers to the stratification of training data. Data stratification and subsequent model training was employed as a simple, straightforward technique to include prior knowledge. We have proposed two ways of how to use data stratification to "conquer" brain tumor segmentation: First, the targeted application of a specialized model (HGG model) to the respective data (HGG test case). Second, the utilization of disagreement among specialized models' outputs and ground truth segmentations to identify outliers and possible latent factors hampering generalization.

Implicitly adding prior information to the models by stratifying the data by tumor type (HGG and LGG) seems to be beneficial for the segmentation of the tumor core for high-grade glioma. Yet, the LGG-only model, which was trained with fewer samples ($N = 75$) compared to the baseline model ($N = 285$), showed no statistically significant deterioration of the segmentation performance. A statistically significant

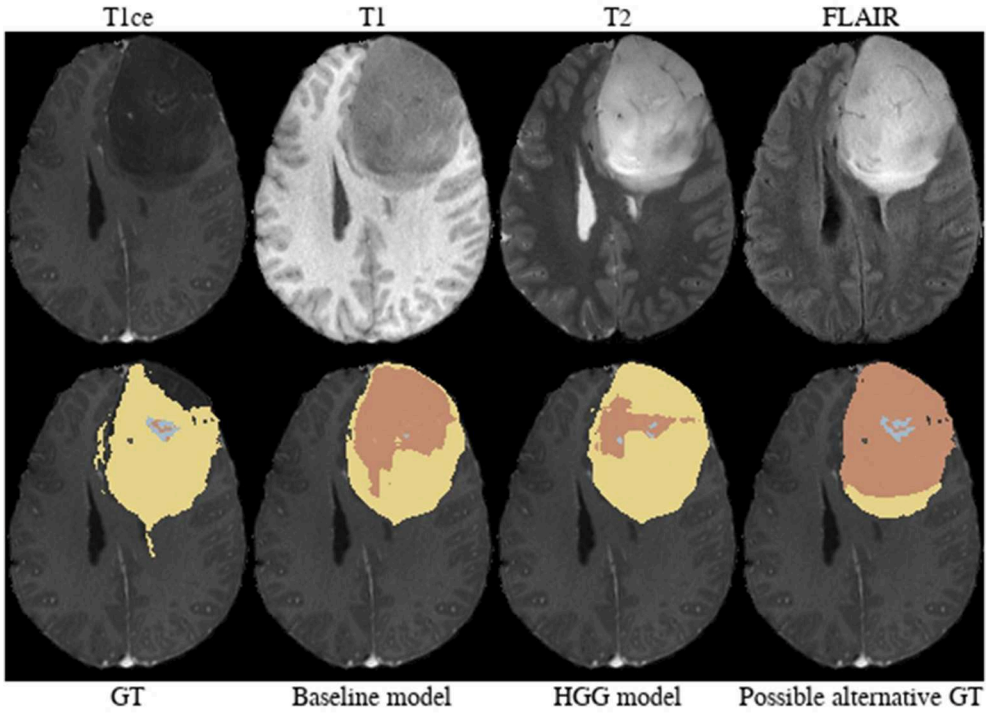


FIGURE 6 | Brats18_CBICA_AXJ_1.

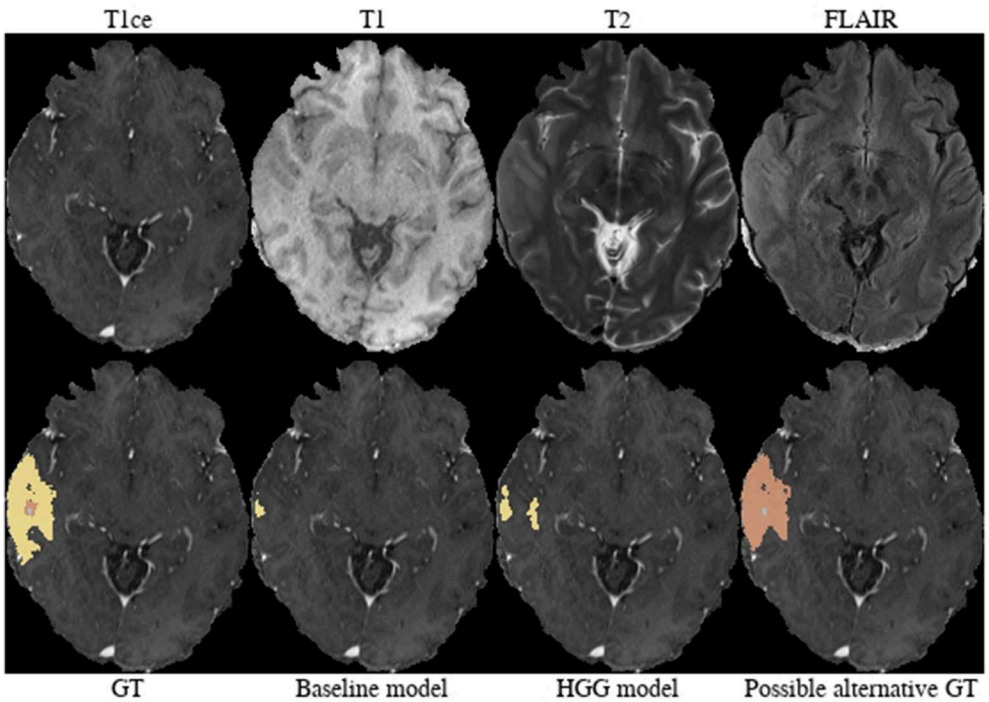


FIGURE 7 | Brats18_CBICA_BHB_1.

improvement in 64.9% of the subjects for the tumor core is accompanied by a non-significant improvement of 54.6% for contrast enhancing and non-significant decrease (only 48.8% better-ranked subjects) for the whole tumor. It has been shown in multiple studies (Asari et al., 1994; Wiestler et al., 2016; Hsieh et al., 2017) that HGG and LGG tend to exhibit different

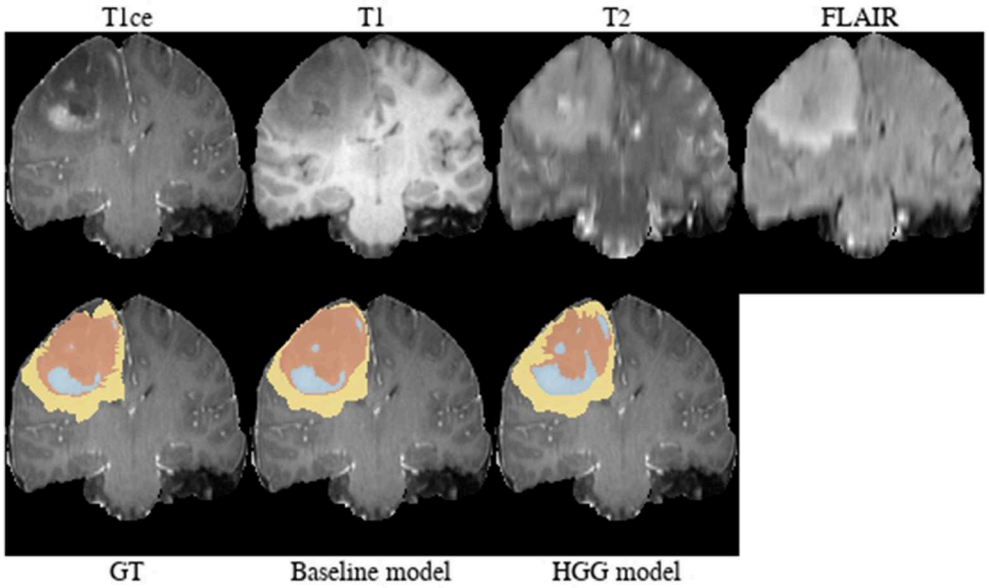


FIGURE 8 | Brats18_TCIA01_221_1.

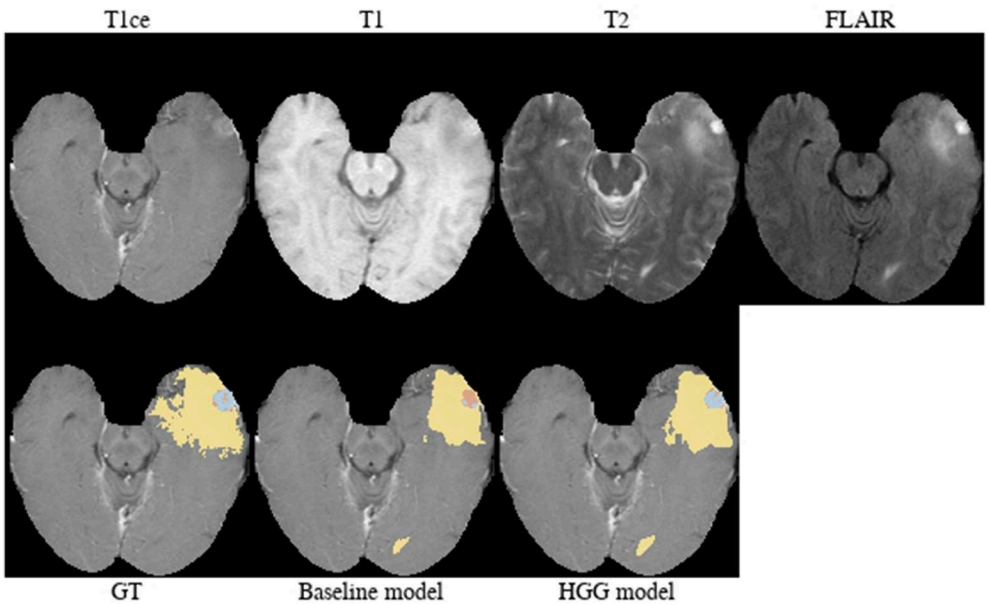


FIGURE 9 | Brats18_TCIA01_425_1.

qualitative and quantitative imaging features in structural MRI, involving heterogeneity of contrast enhancement, cystic components, intratumoral hemorrhage, and necrosis, which in context of tumor segmentation affect the definition of the tumor core greatly. Therefore, the stratification of the training data into HGG and LGG yields subsets with more homogeneous and consistent definitions of the tumor core. However, we presented also exceptions [e.g., BRATS18_2013_21_1 (**Figure 4**) in section 3.2] which actually profit from training data of opposite tumor grade.

In addition to improving segmentation performance, deep learning models trained on stratified data can be used to drive exploration of the training data. In section 3.2 we demonstrated that the disagreement between such models in relation to the ground truth data can assist in the identification of latent factors (e.g., imaging phenotypes) which may pose significant challenges in a deep learning model's capability to generalize across the complete problem domain. We argue that especially in a pathology as complex as brain cancer, the identification of such latent factors and their proper treatment in a deep learning

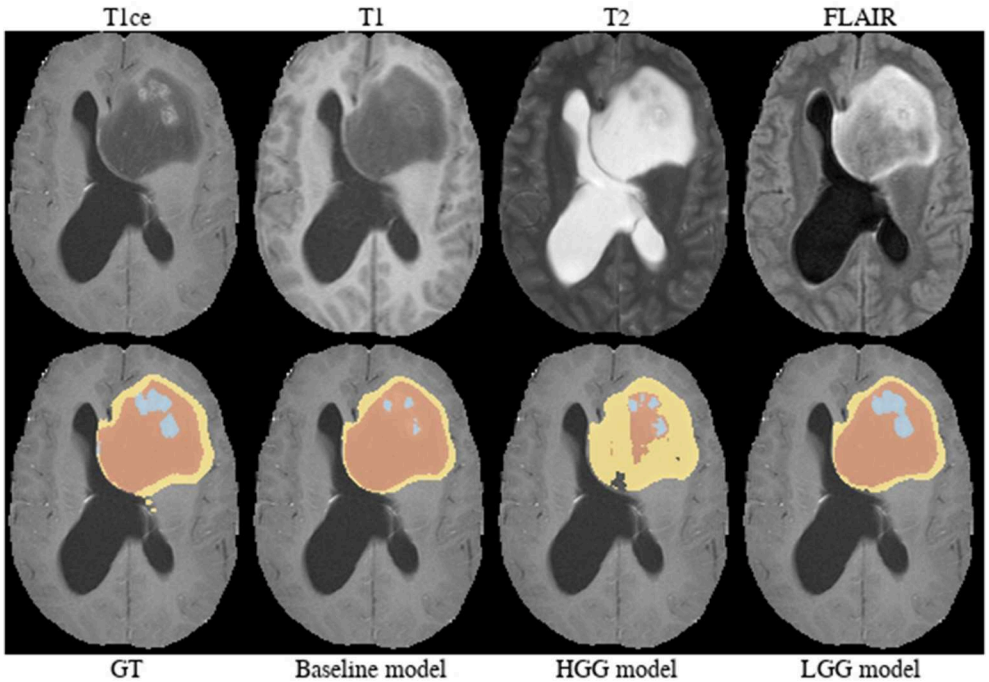


FIGURE 10 | Brats18_TCIA05_444_1.

model is of utmost importance to guarantee robust segmentation performance that satisfies clinical needs. In section 4.1 we provide propositions on how latent factors such as the tumor type could be treated in deep learning segmentation models beyond simple data stratification.

Our results demonstrated the potential of summary statistics (e.g., mean or median) to obfuscate significant differences between distributions of segmentation performance measures (e.g., Dice coefficient). These significant differences can be revealed through the calculation of a case-based ranking. Furthermore, case-based ranking enables the straightforward application of nonparametric statistics to detect significant differences with the advantage of more limited assumptions regarding the distribution of the data when compared to parametric statistics, and robustness to outliers. Case-based ranking also follows the narrative of precision medicine in which the identification of subpopulations of patients, who benefit from a medical intervention, based on experimental observations is central. It enables a more fine-grained analysis on the level of the patient and potentially an identification of patient subpopulations relevant for the task at hand.

Previously, Pereira et al. (2016) trained on data stratified into HGG and LGG. They employed two different Convolutional Neural Network architectures for patch-wise segmentation of HGG and LGG. In contrast, we hypothesized and demonstrated that a mere stratification of the training data into HGG and LGG without any changes to architectures or hyperparameters can lead to improved segmentation performance. Furthermore, their focus was on an ablation study of methodological components with respect to their two grade-specific architectures and their results were based on the BraTS 2013 Leaderboard dataset (21 HGG, 4 LGG cases) and BraTS 2013 Challenge dataset (10 HGG cases).

4.1. Outlook

With the rise of precision medicine and tailored therapies, the consideration of patient-specific information (e.g., genetics) becomes ubiquitous (Giardino et al., 2017). Leveraging data from multiple sources remains a challenge for the next generation imaging technologies (Kim et al., 2016), potentially requiring to rethink the *one size fits all* concept. For automatic brain tumor segmentation, various architectural and conceptual changes are imaginable beyond simple data stratification strategies.

By completely separating the data, each of the individual models has fewer data available for training, although with the benefit of a less heterogeneous domain (only one tumor type). Instead of implicitly adding the prior information of the tumor to the data by stratification, an alternative approach could be to explicitly add this information as input to the network. Particularly the first layers of the network might be less susceptible to the tumor type as filters for representation learning could share commonalities between both domains. By adding the information directly to the input layer or injecting it into the latent feature space might allow the network to intrinsically adapt the segmentation output according to the given tumor type.

A different approach would be to regard the problem of segmenting high-grade and low-grade glioma as a multiple-source adaptation problem. In this setting, the goal is to effectively combine base learners trained on

multiple source domains in order to perform a prediction on a target domain, which can be any mixture of the source domains. In our case, the source domains would be subclasses of gliomas: either high-grade and low-grade data, or potentially a more fine-grained subdivision (e.g., WHO grade or classification). The target domain constitutes of a mix of different glioma cases. Recently, a number of theoretic and algorithmic contributions were made in the area of multiple-source adaptation (Hoffman et al., 2018; Zhao et al., 2018), which could be applied in the scenario of learning from multiple disease entities such as brain tumor types or grades.

The clinical importance of brain tumor segmentation for quantitative image analysis will only grow in the near future. Recently, various segmentation methods have been proposed which are capable of accurately delineating brain tumor compartments longitudinally (Weizman et al., 2014; Meier et al., 2016), perform assessment of treatment response (Huber et al., 2017; Kickingereder et al., 2019), are used for the purpose of radiomic analysis (Bakas et al., 2017c), and for performing planning of radiation therapy (Sharp et al., 2014; Herrmann et al., 2018; Agn et al., 2019; Lipkova et al., 2019). It is, therefore, necessary to provide automatic segmentation methods which are capable of robustly generalizing across different types or grades of brain tumors. Our methodology of training deep learning models on stratified training data is a straightforward approach to potentially improve the segmentation performance of already existing learning-based methods with regards to different tumor types.

In the light of our results and the trend toward precision medicine, we encourage challenge organizers to make information on the tumor type or grade available as additional input data, allowing teams to incorporate such prior information into their models.

4.2. Limitations

The evaluation is based solely on the BraTS training dataset (using cross-validation). Results for the official validation set are unknown since the required tumor type is not available for these data. Indeed we acknowledge that the tumor grade is usually not yet available on the first admission. However, we think automatic segmentation models will probably be employed first for retrospective studies, to assess the extent of resection in patients undergoing surgery (Meier et al., 2017), or to assess tumor progression postoperatively (Kickingereder et al., 2019) where tumor grades are usually known. First attempts have been made to classify tumor grades from MRI (Decuyper and Van Holen, 2019), which would allow identification of the correct model from imaging only. Alternatively, one might run such a segmentation algorithm twice: first for a rough identification of the tumor compartments and based on the result (e.g., presence of CE, ratio of compartment volumes, or manual review of the intermediate results by an expert) apply the specific model to get a refined segmentation.

The benefit of stratifying the training data has been shown with the model ranked third in the BraTS 2018 challenge (McKinley et al., 2019a). This particular model was chosen, as it was a top-ranked method in the most recent BraTS challenge (2018) that achieved its results using only a standard GPU and data from the BRATS challenge. The method ranked first (Myronenko, 2019) depended on a GPU with 32 GB of memory (to which most research groups do not have access), while the second-ranked method (Isensee et al., 2018) was co-trained with additional data (not including information about tumor grades). To what extent the proposed approach generalizes to other architectures remains an open question. Other models might suffer more from the reduction of training samples due to the stratification. The proposed architecture is known to be robust to fewer training samples (McKinley et al., 2019b).

5. CONCLUSION

Implicitly adding prior knowledge by dividing data into distinct domains can improve the performance of deep learning-based segmentation methods and compensate for the smaller number of samples available for training a model. The tumor grade has shown to be an important latent factor in the segmentation of gliomas. Comparing the performance of models by case-based ranking statistics may reveal significant differences that are otherwise concealed in summary statistics such as the mean Dice coefficient.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://braintumorsegmentation.org.

ETHICS STATEMENT

The study is based on a publicly available dataset from the Multimodal Brain Tumor Segmentation Challenge 2018 (BraTS, http://braintumorsegmentation.org/).

AUTHOR CONTRIBUTIONS

MReb, RM, RMc, RW, and MRey: design of experiments. MReb: perform experiments, data analysis. UK and RM: analysis of selected cases. MReb, RM, and RMc: results interpretation. MReb and RM: manuscript drafting, RMc: manuscript revision. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work was supported by the Swiss Personalized Health Network (SPHN, project number 2018DRI10), the Swiss National Science Foundation (grant number 169607), and the Swiss Cancer League (grant number KFS-3979-08-2016).

ACKNOWLEDGMENTS

Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01182/full#supplementary-material

REFERENCES

Agn, M., Af Rosenschöld, P. M., Puonti, O., Lundemann, M. J., Mancini, L., Papadaki, A., et al. (2019). A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Med. Image Anal.* 54, 220–237. doi: 10.1016/j.media.2019.03.005

Asari, S., Makabe, T., Katayama, S., Itoh, T., Tsuchida, S., and Ohmoto, T. (1994). Assessment of the pathological grade of astrocytic gliomas using an MRI score. *Neuroradiology* 36, 308–310. doi: 10.1007/BF00593267

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). Segmentation labels for the pre-operative scans of the TCGA-GBM collection. *The Cancer Imaging Archive*. doi: 10.7937/k9/tcia.2017.klxwjj1q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels for the pre-operative scans of the TCGA-LGG collection. *The Cancer Imaging Archive*. doi: 10.7937/k9/tcia.2017.gjq7r0ef

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *CoRR* abs/1811.02629.

Carrillo, J., Lai, A., Nghiemphu, P., Kim, H., Phillips, H., Kharbanda, S., et al. (2012). Relationship between tumor enhancement, edema, IDH1 mutational status, MGMT promoter methylation, and survival in glioblastoma. *Am. J. Neuroradiol.* 33, 1349–1355. doi: 10.3174/ajnr.A2950

Decuyper, M., and Van Holen, R. (2019). "Fully automatic binary glioma grading based on pre-therapy MRI using 3D convolutional neural networks," *Presented at the International Conference on Medical Imaging with Deep Learning, MIDL 2019* (London).

Ellingson, B. M., Bendszus, M., Boxerman, J., Barboriak, D., Erickson, B. J., Smits, M., et al. (2015). Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro Oncol.* 17, 1188–1198. doi: 10.1093/neuonc/nov095

Ellingson, B. M., Cloughesy, T. F., Pope, W. B., Zaw, T. M., Phillips, H., Lalezari, S., et al. (2012). Anatomic localization of o6-methylguanine DNA methyltransferase (MGMT) promoter methylated and unmethylated tumors: a radiographic study in 358 *de novo* human glioblastomas. *NeuroImage* 59, 908–916. doi: 10.1016/j.neuroimage.2011.09.076

Essig, M., Anzalone, N., Combs, S., Dörfler, A., Lee, S.-K., Picozzi, P., et al. (2012). MR imaging of neoplastic central nervous system lesions: review and recommendations for current practice. *Am. J. Neuroradiol.* 33, 803–817. doi: 10.3174/ajnr.A2640

Gevaert, O., Mitchell, L. A., Achrol, A. S., Xu, J., Echegaray, S., Steinberg, G. K., et al. (2014). Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology* 273, 168–174. doi: 10.1148/radiol.14131731

Giardino, A., Gupta, S., Olson, E., Sepulveda, K., Lenchik, L., Ivanidze, J., et al. (2017). Role of imaging in the era of precision medicine. *Acad. Radiol.* 24, 639–649. doi: 10.1016/j.acra.2016.11.021

Grier, J. T. (2006). Low-grade gliomas in adults. *Oncologist* 11, 681–693. doi: 10.1634/theoncologist.11-6-681

Grossmann, P., Gutman, D. A., Dunn, W. D., Holder, C. A., and Aerts, H. J. W. L. (2016). Imaging-genomics reveals driving pathways of MRI derived volumetric tumor phenotype features in glioblastoma. *BMC Cancer* 16:611. doi: 10.1186/s12885-016-2659-5

Herrmann, E., Ermiş, E., Meier, R., Blatti-Moreno, M., Knecht, U., Aebersold, D., et al. (2018). Fully automated segmentation of the brain resection cavity for radiation target volume definition in glioblastoma patients. *Int. J. Radiat. Oncol. Biol. Phys.* 102:S194. doi: 10.1016/j.ijrobp.2018.07.087

Hoffman, J., Mohri, M., and Zhang, N. (2018). "Algorithms and theory for multiple-source adaptation," in *Advances in Neural Information Processing Systems 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC: Curran Associates, Inc.), 8246–8256.

Hsieh, K. L.-C., Tsai, R.-J., Teng, Y.-C., and Lo, C.-M. (2017). Effect of a computer-aided diagnosis system on radiologists' performance in grading gliomas with MRI. *PLoS ONE* 12:e0171342. doi: 10.1371/journal.pone.0171342

Huber, T., Alber, G., Bette, S., Kaesmacher, J., Boeckh-Behrens, T., Gempt, J., et al. (2017). Progressive disease in glioblastoma: benefits and limitations of semi-automated volumetry. *PLoS ONE* 12:e0173112. doi: 10.1371/journal.pone.0173112

Hygino da Cruz, L., Rodriguez, I., Domingues, R., Gasparetto, E., and Sorensen, A. (2011). Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment glioma. *Am. J. Neuroradiol.* 32, 1978–1985. doi: 10.3174/ajnr.A2397

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 234–244.

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., et al. (2019). Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 20, 728–740. doi: 10.1016/S1470-2045(19)30098-1

Kim, M. M., Parolia, A., Dunphy, M. P., and Venneti, S. (2016). Non-invasive metabolic imaging of brain tumours in the era of precision medicine. *Nat. Rev. Clin. Oncol.* 13, 725–739. doi: 10.1038/nrclinonc.2016.108

Lai, A., Eskin, A., Ellingson, B. M., Phillips, H. S., Nghiemphu, P. L., Chowdhury, R., et al. (2013). Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images. *Neuro Oncol.* 15, 626–634. doi: 10.1093/neuonc/not008

Lasocki, A., and Gaillard, F. (2019). Non-contrast-enhancing tumor: a new frontier in glioblastoma research. *Am. J. Neuroradiol.* 40, 758–765. doi: 10.3174/ajnr.A6025

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lipkova, J., Angelikopoulos, P., Wu, S., Alberts, E., Wiestler, B., Diehl, C., et al. (2019). Personalized radiotherapy design for glioblastoma: integrating mathematical tumor models, multimodal scans and bayesian inference. *IEEE Trans. Med. Imaging* 38, 1875–1884. doi: 10.1109/TMI.2019.2902044

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* 9:5217. doi: 10.1038/s41467-018-07619-7

McKinley, R., Meier, R., and Wiest, R. (2019a). "Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 456–465.

McKinley, R., Rebsamen, M., Meier, R., Reyes, M., Rummel, C., and Wiest, R. (2019b). Few-shot brain segmentation from weakly labeled data with deep heteroscedastic multi-task networks. *arXiv [preprint]. arXiv:1904.02436*.

Meier, R., Knecht, U., Loosli, T., Bauer, S., Slotboom, J., Wiest, R., et al. (2016). Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Sci. Rep.* 6:23376. doi: 10.1038/srep23376

Meier, R., Porz, N., Knecht, U., Loosli, T., Schucht, P., Beck, J., et al. (2017). Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma. *J. Neurosurg.* 127, 798–806. doi: 10.3171/2016.9.JNS16146

Meier, R., Rebsamen, M., Knecht, U., Reyes, M., Wiest, R., and McKinley, R. (2019). "Stratify or inject: two simple training strategies to improve brain tumor segmentation," in *Presented at the International Conference on Medical Imaging with Deep Learning, MIDL 2019* (London).

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Mullins, M. E., Barest, G. D., Schaefer, P. W., Hochberg, F. H., Gonzalez, R. G., and Lev, M. H. (2005). Radiation necrosis versus glioma recurrence: conventional MR imaging clues to diagnosis. *Am. J. Neuroradiol.* 26, 1967–1972.

Myronenko, A. (2019). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 311–320.

Ostrom, Q. T., Bauchet, L., Davis, F. G., Deltour, I., Fisher, J. L., Langer, C. E., et al. (2014). The epidemiology of glioma in adults: a state of the science review. *Neuro Oncol.* 16, 896–913. doi: 10.1093/neuonc/nou087

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* 35, 1240–1251. doi: 10.1109/TMI.2016.2538465

Pierallini, A., Bonamini, M., Bozzao, A., Pantano, P., Stefano, D. D., Ferone, E., et al. (1997). Supratentorial diffuse astrocytic tumours: proposal of an MRI classification. *Eur. Radiol.* 7, 395–399. doi: 10.1007/s003300050173

Pope, W. B., Sayre, J., Perlina, A., Villablanca, J. P., Mischel, P. S., and Cloughesy, T. F. (2005). MR imaging correlates of survival in patients with high-grade gliomas. *Am. J. Neuroradiol.* 26, 2466–2474.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sharp, G., Fritscher, K. D., Pekar, V., Peroni, M., Shusharina, N., Veeraraghavan, H., et al. (2014). Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med. Phys.* 41:050902. doi: 10.1118/1.4871620

Sottoriva, A., Spiteri, I., Piccirillo, S. G. M., Touloumis, A., Collins, V. P., Marioni, J. C., et al. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4009–4014. doi: 10.1073/pnas.1219747110

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New Engl. J. Med.* 352, 987–996. doi: 10.1056/NEJMoa043330

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE International Conference on Computer Vision, ICCV 2017* (Venice), 843–852.

Treiber, J. M., Steed, T. C., Brandel, M. G., Patel, K. S., Dale, A. M., Carter, B. S., et al. (2018). Molecular physiology of contrast enhancement in glioblastomas: an analysis of the cancer imaging archive (TCIA). *J. Clin. Neurosci.* 55, 86–92. doi: 10.1016/j.jocn.2018.06.018

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Weizman, L., Sira, L. B., Joskowicz, L., Rubin, D. L., Yeom, K. W., Constantini, S., et al. (2014). Semiautomatic segmentation and follow-up of multicomponent low-grade tumors in longitudinal brain MRI studies. *Med. Phys.* 41:052303. doi: 10.1118/1.4871040

Wen, P. Y., MacDonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., et al. (2010). Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J. Clin. Oncol.* 28, 1963–1972. doi: 10.1200/JCO.2009.26.3541

Wiestler, B., Kluge, A., Lukas, M., Gempt, J., Ringel, F., Schlegel, J., et al. (2016). Multiparametric MRI-based differentiation of WHO grade II/III glioma and WHO grade IV glioblastoma. *Sci. Rep.* 6:35142. doi: 10.1038/srep35142

Zhao, H., Zhang, S., Wu, G., Moura, J. M. F., Costeira, J. P., and Gordon, G. J. (2018). "Adversarial multiple source domain adaptation," in *Advances in Neural Information Processing Systems 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc.), 8559–8570.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

BRIEF RESEARCH REPORT
published: 08 November 2019
doi: 10.3389/fncom.2019.00073

Robustness of Radiomics for Survival Prediction of Brain Tumor Patients Depending on Resection Status

Leon Weninger, Christoph Haarburger and Dorit Merhof*

Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany

Prediction of overall survival based on multimodal MRI of brain tumor patients is a difficult problem. Although survival also depends on factors that cannot be assessed via preoperative MRI such as surgical outcome, encouraging results for MRI-based survival analysis have been published for different datasets. We assess if and how established radiomic approaches as well as novel methods can predict overall survival of brain tumor patients on the BraTS challenge dataset. This dataset consists of multimodal preoperative images of 211 glioblastoma patients from several institutions with reported resection status and known survival. In the official challenge setting, only patients with a reported gross total resection (GTR) are taken into account. We therefore evaluated previously published methods as well as different machine learning approaches on the BraTS dataset. For different types of resection status, these approaches are compared to a baseline, a linear regression on patient age only. This naive approach won the 3rd place out of 26 participants in the BraTS survival prediction challenge 2018. Previously published radiomic signatures show significant correlations and predictiveness to patient survival for patients with a reported subtotal resection. However, for patients with reported GTR, none of the evaluated approaches was able to outperform the age-only baseline in a cross-validation setting, explaining the poor performance of approaches based on radiomics in the BraTS challenge 2018.

Keywords: BraTS 2018, survival prediction, radiomics, brain tumor, machine learning, feature selection

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Hamed Akbari,
University of Pennsylvania,
United States
Zhi-Cheng Li,
Shenzhen Institutes of Advanced Technology (CAS), China

****Correspondence:***
Leon Weninger
leon.weninger@lfb.rwth-aachen.de

Received: *25 April 2019*
Accepted: *09 October 2019*
Published: *08 November 2019*

Citation:
Weninger L, Haarburger C and Merhof D (2019) Robustness of Radiomics for Survival Prediction of Brain Tumor Patients Depending on Resection Status. Front. Comput. Neurosci. 13:73. doi: 10.3389/fncom.2019.00073

1. INTRODUCTION

The high-grade glioma, a subtype of brain tumors, is one of the most aggressive and dangerous diseases worldwide. For the US, a 5-year survival rate of glioblastoma patients of only 5.6% was reported for 2000–2015 (Ostrom et al., 2018). Automatic analysis of these tumors is challenging, as their shape, location and extent can differ substantially. Since 2012, the BraTS challenge (Menze et al., 2015) is held annually to allow an unbiased comparison of different segmentation algorithms. Since 2017, an overall survival (OS) prediction task is included to assess whether quantitative image features based on these segmentations can provide further clinical insight. In the OS task, patients need to be classified in *long-survivors* (OS>15 months), *short-survivors* (OS <10 months), and *mid-survivors* (10 months <OS <15 months). While data is provided for patients with different resection status, the official evaluation is carried out only on patients with a reported gross total resection (GTR). A total of 41 teams took part in this survival prediction task in 2017 and 2018.

Using the age as sole feature with a linear regressor, we achieved an accuracy of 0.56 (n = 77) on the test set in the BraTS challenge 2018. In comparison, the first placed approaches of 2017 (Shboul et al., 2018) and 2018 (Feng et al., 2019) achieved accuracies of around 0.58 and 0.62, respectively (Bakas et al., 2018b). Shboul et al. relied on automatic radiomic feature extraction combined with a Random Forest Regressor (RFR), while Feng et al. used geometric features in combination with a linear model. The developers of other top performing algorithms chose similar strategies of combining either hand-selected or automated radiomic features with a supervised machine learning algorithm: Radiomic feature extraction was used in combination with an RFR (Sun et al., 2019) or a Multilayer Perceptron (MLP) (Baid et al., 2019). Geometric features only were used with an MLP (Jungo et al., 2018), and finally atlas locations together with relative tumor sizes and an RFR were also employed (Puybareau et al., 2019). These teams achieved accuracies between 0.55 and 0.6. Further submitted approaches ranged from deep learning algorithms to radiomic feature analysis to handcrafted feature engineering, that achieved accuracies between 0.15 and 0.55. As three classes were equally subdivided, a random choice would result in an accuracy of 0.33.

On other brain tumor datasets, encouraging results for OS prediction have been published. A successful radiomic-based brain tumor patient OS and progression-free survival prediction on a private dataset comprising 119 patients was described by Kickingereder et al. (2016). Positive findings with data-mining algorithms have also been reported when including Diffusion-MRI and relative cerebral blood volume data (Zacharaki et al., 2012) or Perfusion-MRI data (Jain et al., 2014) next to the MR-sequences used in the BraTS dataset. Deep learning based OS prediction has been successfully used on another, smaller (n = 93) private dataset (Nie et al., 2019). However, as the BraTS summary (Bakas et al., 2018b) indicates, deep learning techniques performed rather poorly on the open-access data. Quantitatively comparing deep learning to classical regression on radiomic features for OS on the BraTS data was also carried out by Suter et al. (2019). They concluded that radiomic feature are better suited, as features extracted from deep learning networks seemed to be unstable for this task.

Radiomic feature extraction describes the process of automatically computing a variety of quantitative image features. By quantifying lesions, radiomics can not only be used for prognosis, but can also help increase precision in diagnosis. For example, radiomics has been successfully used to distinguish between high- and low-grade glioma (Cho et al., 2018) on the BraTS dataset. An overview of radiomics and its applications is given by Rizzo et al. (2018). For brain tumor analysis in particular, a review of radiomics-based techniques for quantitative imaging is given by Zhou et al. (2018).

Radiomic features combined with a machine learning model is thus a natural choice for OS prediction. We initially evaluated different radiomics-based machine learning techniques for the BraTS challenge, too. However, when thoroughly validating the results, all considered approaches could not outperform a linear regressor based on the patients age only. We thus decided to submit an age-only linear regressor (Weninger et al., 2019), and won the third place in the BraTS challenge 2018.

In this paper, we analyze different radiomic-based approaches to survival prediction on the BraTS data. To be independent of segmentation inaccuracies, we only use the BraTS training data for all experiments. For this data, groundtruth segmentations are publicly available, approved by experts and reviewed by a single board-certified neuro-radiologist (Bakas et al., 2017c). The data can be subdivided by resection status into patients with reported GTR, subtotal resection (STR) and patients with unavailable resection status (NA). The official evaluation was carried out only on the GTR subset. First, we re-evaluate previously published radiomic signatures on the different resection status subsets. We show that these methods are predictive for OS on the STR subset. Second, different machine learning tools are evaluated on the radiomic feature set. Third, as the number of extracted radiomic features is very large and important features could remain undetected, two different feature reduction methods are assessed.

For the patients with GTR, neither previously published methods, nor different machine learning models, nor unsupervised feature reduction techniques could establish a robust signature for patient survival prediction. Finally, the importance of thoroughly assessing the robustness of radiomic markers is discussed, and ideas on how to improve survival prediction based on MRI images even after tumor resection are provided.

2. MATERIALS

2.1. Dataset

In our evaluation, we discard the BraTS test- and validation datasets, as no groundtruth segmentations and no OS information are available, and use only the training dataset. All subjects of the BraTS 2018 dataset are included in the BraTS 2019 dataset; thus, the analysis is focused on the larger BraTS 2019 dataset. The BraTS survival data training dataset consists of data from 211 brain tumor patients from different institutions. For each patient, the following data is available:

- 4 MRI acquisitions: T1, T1 post contrast agent (T1CE), T2 and T2-FLAIR. All are resampled to an isotropic resolution of $1 \times 1 \times 1\text{mm}^3$, co-registered and skull stripped.
- Segmentation map: Edema (ED), enhancing tumor (ET), and non-enhancing / necrotic tumor core (NEC).
- The age of the patient.
- Resection status.

The resection status is either reported as GTR, subtotal resection (STR), or unknown (NA). For a few subjects (n = 21), the resection status was given as STR in the BraTS 2018 dataset, but omitted for the 2019 dataset. These statuses were re-entered into the dataset. Next, two patients were reported as still alive. Their overall survival in the database was set to the maximum survival time in the dataset, 1,767 days.

2.2. Cohort Study

Most data are provided either by the Center for Biomedical Image Computing and Analytics from University of Pennsylvania (CBICA, n = 128) or by The Cancer Imaging Archive (TCIA, n = 76) (Bakas et al., 2017a,b). A small amount of the data (n = 7)

originates from other sources. All subjects have a pathologically confirmed diagnosis of primary *de novo* glioblastoma (Bakas et al., 2018b). Nevertheless, as population or differences in treatment could influence clinical outcome, an overview over differences and similarities of the different provenances is given.

For all TCIA subjects, the resection status is unknown. In contrast, 94 of the 101 subjects with GTR as well as all subjects with STR originate from one institution, CBICA. In the dataset, there are no statistically significant differences between age or survival for the different data provenances or the different types of resection status (ANOVA: $p > 0.05$). However, the relative brain tumor volume, as determined as tumor volume divided by brain volume, is significantly smaller in the TCIA data than in the CBICA data ($p < 0.0001$). Between the resection status STR and GTR, in contrast, there is no significant relative brain tumor volume difference (**Figure 1**).

3. METHODS

Our OS prediction pipeline can be divided into the following substeps: (1) Image preprocessing, (2) extraction of radiomic features, (3) unsupervised feature reduction, and (4) statistical inference and out-of-sample prediction. These major substeps of the pipeline are visualized in **Figure 2**. For the BraTS challenge, only out-of-sample prediction is necessary. In order to determine whether radiologic features are appropriate for the given problem, we supplement out-of-sample prediction with classical hypothesis testing.

3.1. Image Preprocessing

The data was acquired with various MRI scanners and different clinical protocols. In consequence, absolute image intensities, and, subsequently, radiomic features, can be strongly influenced. This was counteracted with a bias-filed correction and subsequent normalization of the images. First, the ANTs N3 (Tustison et al., 2010) bias-field correction was applied to all images, removing local differences in image intensities. Second, in order to harmonize the MRI acquisitions from different institutions, all images were normalized with z-score normalization to zero mean and unit variance.

Histogram equalization was considered as alternative normalization technique, but discarded as it did not improve the results. This could be due to the properties of tumor tissue in MRI images: Parts of the brain tumor are often the brightest or darkest area in the acquisitions, while occupying only a small proportion of the brain. The contrast-enhancing part is especially bright in T1CE acquisitions while covering just a small single-digit percentage of the brain volume. Histogram equalization or other nonlinear brightness adaptation techniques will thus shrink the contrast for these outlier points, actually leading to less contrast in the examined regions. For a comparison of the results using histogram equalization, all evaluations relying not only on tumor shape and/or age were repeated with histogram equalization instead of z-score normalization. The results can be found in the **Supplementary Materials**.

3.2. Feature Extraction

Using the package *PyRadiomics* (van Griethuysen et al., 2017), shape features were extracted from the provided segmentation masks, and image intensity and texture features were extracted from the four different image modalities for each segmentation mask. Image intensity and texture features were calculated for the original image and on wavelet decomposed images. In total, the following features were extracted:

Shape features comprise volume, surface area, sphericity, maximum diameter, elongation, axis lengths and flatness. These were extracted for the different tumor classes, resulting in 42 features.

Gray-level features include gray-level co-occurrence (glcm), gray-level run length (glrlm), gray-level dependence matrix (gldm), gray-level size zone, and neighboring gray tone difference features. As these were extracted for the original and wavelet transformed images and four image modalities, this resulted in 7,884 features.

Image intensity statistics consists of features such as minimum, maximum, mean, median, percentiles, standard deviation, skewness, kurtosis, and uniformity. In combination with different modalities and filters, 1,944 features resulted from this category.

Combined with the age, a total of 9871 features were obtained. In contrast, the total number of observations was 211—the number of variables p is much bigger than the number of samples n. Such a setting is actually common for pattern-learning methods in neuroscience (Bzdok, 2017), and is referred as *wide data*, in contrast to *long data* where the number of samples is bigger than the number of variables. Using such wide data directly for inference often leads to non-robust results and to overfitting on the training set. Consequently, before inference the number of features needs to be reduced as much as possible while maintaining the characteristics of the data.

3.3. Preselection of Features

Radiomic features are typically redundant (Rizzo et al., 2018), i.e., they are multicollinear. Different techniques exist to reduce the number of features and thus the multicollinearity. For the present problem, a subset of features should be kept after feature reduction. In contrast to synthetic features obtained by a PCA, a feature selection method offers more interpretable results. Further, in order to use the complete BraTS training dataset, the method should be unsupervised. With an unsupervised method, the complete BraTS OS training data ($n = 211$) can be used for feature selection, as features of preoperative images should be independent of resection status. In contrast, for this study, a supervised method could only be done on the specific resection status subset (GTR: $n = 101$). As splitting into train- and test set would further be necessary, an even smaller number of examples could be employed for feature selection.

Thus, a method relying on correlation matrix clustering and Variance-Inflation-Feature (VIF) iterative reduction (James et al., 2014) was chosen as the most appropriate. As a first step to reduce multicollinearity, single redundant features were discarded. For this purpose, each feature was linearly

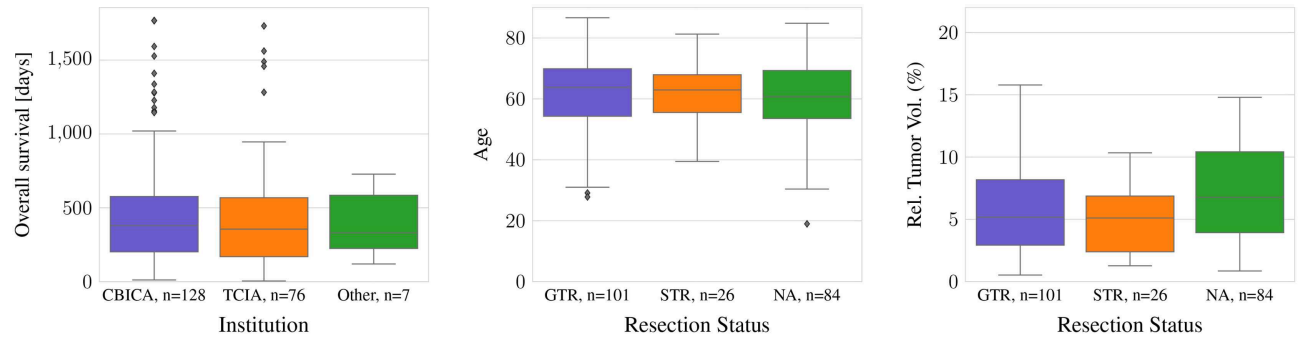


FIGURE 1 | Difference in age and tumor-to-brain volume by resection status, and survival by provenance.

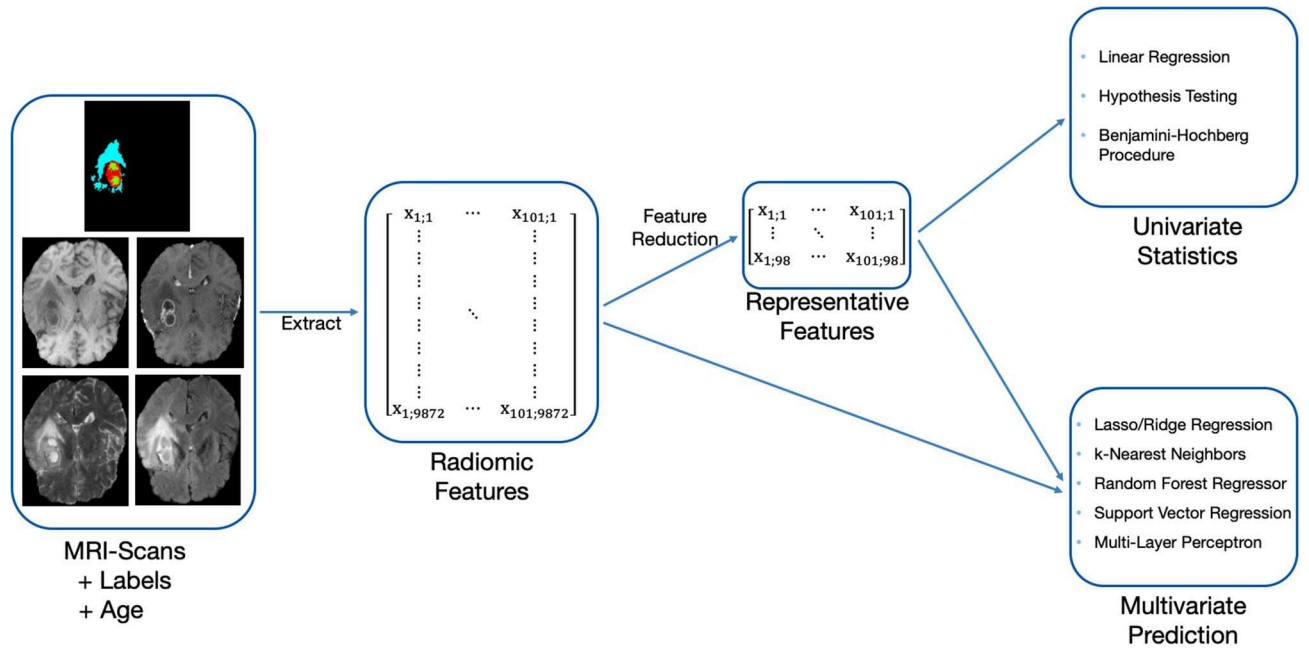


FIGURE 2 | Methodology used to evaluate the predictiveness of radiomic features for OS with the available acquisitions and labels.

regressed against every other single feature, thus obtaining the coefficient of determination r^2 and creating an r^2 correlation matrix. This matrix was then reordered using a hierarchical clustering algorithm. For this, we relied on the Voor Hees Algorithm (Voorhees, 1986) implemented in SciPy (Jones et al., 2001) for linkage, and Euclidian distances between rows or columns of the correlation matrix. A visual impression of the obtained clustered correlation matrix is given in **Figure 3**.

As proposed by Gillies et al. (2016), representative features can be chosen from each cluster to reduce redundant elements. For this, areas of high correlation ($R^2 > 0.95$) were reduced to the element with the highest inter-patient variability. Using this method, only features having a pairwise collinear correlation can be identified and omitted. Multicollinearity, i.e., highly related associations between more than two features, is not taken into account.

Multicollinear features were excluded in a second step. Those features can be identified by checking the VIF. Iteratively, by removing the feature with the highest VIF, the multicollinearity can be reduced until a predefined threshold is obtained. A maximum VIF of 10 is chosen, as thresholds of either 5 or 10 are recommended for this method (James et al., 2014). The number of features retained with a threshold of 10 should not pose problems to the machine learning models, so we did not consider lower thresholds.

Next to the VIF-based feature preselection method, we evaluated a principal component analysis (PCA) based feature reduction pipeline. One PCA feature reduction was carried out independently for the shape features, gray-level features and image intensity features of the original image. A fourth PCA was performed on all features of wavelet decomposed images. For each analysis, the minimum number of principal

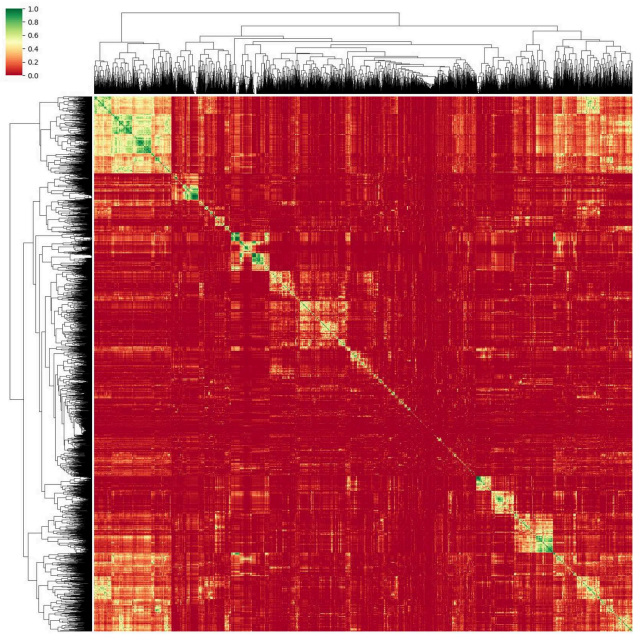


FIGURE 3 | Reordered correlation matrix and obtained dendrogram of the radiomic features obtained from the normalized MRI images. Green designates strongly correlated features while red indicates an r^2 close to zero.

components explaining 95% of variance in the data were kept. The obtained features are finally concatenated, and the predictiveness for survival prediction can be evaluated via machine learning models.

3.4. Statistical Hypothesis Testing on Single Features

Null hypothesis testing with false discovery rate correction on the original dataset is not beneficial, as there are too many correlated features. The subset selected by the VIF feature selection (section 3.3), however, is much smaller and hypothesis tests can now reveal if single features are actually predictive for OS. As multiple radiomic features remained, a false discovery rate correction still needs to be used. We relied on the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), that controls the false discovery rate at a specific level $\alpha = 0.05$.

3.5. Multivariate Prediction

The statistical hypothesis testing can only reveal if single features are significantly predictive for OS. Nonlinear relationships of single predictors to the target variable as well as feature interactions cannot be detected. Different machine learning models that are able to surpass this limitation are available, ranging in complexity from basic linear regressors to complex neural networks.

We evaluated different machine learning models: linear, lasso and ridge regressors, k-nearest neighbors (kNN), random forests regressors (RFR), support vector regressors (SVR), and support vector classifiers (SVC). Furthermore, the Boruta (Kursa and Rudnicki, 2010) feature selection algorithm in combination with one random forest classifier (RFC) as estimator and one for the final prediction was evaluated. The regression models were directly fitted to the survival days, while the classifier can only predict the classes. As classes, the three classes as proposed by the BraTS challenge (*long-survivors* (OS >15 months), *short-survivors* (OS <10 months), and *mid-survivors* (10 months <OS <15 months) were used.

Different radiomic features are represented by absolute values at very different scales. Furthermore, outliers of single features may strongly influence the results. Consequently, the radiomic features were first normalized: The feature median is subtracted, and the features were scaled by the interquartile range, i.e., the range between the 25th quantile and the 75th quantile.

The different machine learning models were first employed on the complete feature set for the different resection status. The same methods were then also tested on the VIF-based feature subset as well as on the PCA reduced feature set, in order to evaluate whether these models could improve robustness on GTR patients.

All machine learning models were implemented with scikit-learn v0.21.2 (Pedregosa et al., 2011) or scikit-learn-contrib using default settings. Next to the methodology presented in this paper, we further evaluated the linear regressor on the age only as submitted during the BraTS challenge 2018, as well as a linear regression on age and the features remaining significant after Benjamini–Hochberg correction.

3.6. Evaluation of Previously Published Methods

Previously reported relationships between radiomic signatures and survival time were evaluated on the BraTS dataset. Gutman et al. (2013) reported that the length of the lesion's major axis and the proportion of contrast-enhanced tumor were negatively correlated with survival on the TCGA glioblastoma dataset. It should be noted that this dataset is included in the BraTS dataset with the resection status NA. It has also been shown that volumetric features of enhancing tumor, non-enhancing tumor core and necrosis, and edema normalized to brain volume are associated with shorter survival time on different independent datasets (Zhang et al., 2014; Macyszyn et al., 2015).

Kickingereder et al. (2016) proposed a supervised principal component analysis of radiomic features for glioblastoma patients. In this study, a set of MRI acquisitions also comprising diffusion and susceptibility-weighted MR imaging was used. Thus, compared to our analysis, the study relied on a different set of radiomic features. Nevertheless, their statistical analysis pipeline with z-score feature normalization and supervised principal component analysis is directly applicable to the features described in section 3.2.

4. RESULTS

First, the predictiveness of state-of-the art methods and machine learning model using radiomic features is evaluated for the different types of resection status in sections 4.1 and 4.2.

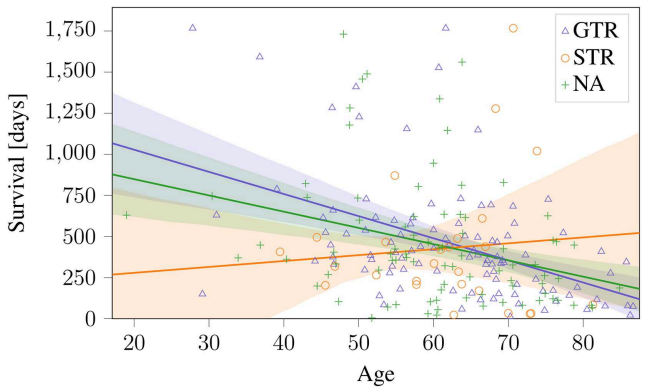


FIGURE 4 | Age-only regression for the different types of resection status on BraTS 2019 training dataset. The obtained regression line is plotted together with the 95% confidence interval. This model is used as baseline against radiomics based approaches.

These sections show that a predictive radiomic signature can be extracted for STR patients. For these patients, radiomics based approaches to survival prediction outperform the age-only approach that can be seen in **Figure 4**. However, on the patients that underwent total resection of the tumor, the findings are different: The established radiomic features as well as all considered machine learning models fail to improve the survival prediction. Regression on age-only, however, is significantly correlated with shorter survival for GTR patients (**Figure 4**).

Then, as the results of different models show a high variability, it is assessed in section 4.3 whether models based on a selected subset of features can lead to more robust results for GTR patients.

4.1. Repeatability of Previous Methods on Dataset

As a first step, previously reported relationships between radiomic signatures and survival time (section 3.6) were evaluated on the BraTS dataset. For evaluation, the dataset was first divided by the three different reported types of resection status. The published radiomic signatures were evaluated on each subset individually.

The two features proposed by Gutman et al. (2013) and the volumetric features of ET and NEC normalized to brain volume proposed by Zhang et al. (2014), Macyszyn et al. (2015) can be seen in **Figure 5**. These findings can be reproduced on the STR, and the same trends can also be seen on the NA subset (Pearson's r: p <0.05). Especially the ET volume and the lesions' major axis achieve a high significance ($p \approx 0.003$) on the STR subset. Of the reported features, the only non-significant relationships are ED volume, that shows a negative, but non-significant (p >0.05) correlation on both subsets, and the ET tumor proportion, that shows a significant negative correlation on the STR subset, but only a non-significant negative correlation on the NA subset. However, on the subset with reported GTR, no correlation can be identified for any feature.

Next, the statistical analysis pipeline for radiomic features of glioblastoma patients proposed by Kickingereder et al. (2016) was applied to the different resection status subsets. In the original publication, MRI acquisitions that are not available in the BraTS data (e.g., diffusion MRI) and slightly different radiomic features were used. Nevertheless, the proposed z-score feature normalization and supervised principal component analysis is directly applicable to the present dataset, and can give a good baseline model. Using the proposed model parameters, the analysis was repeated on the radiomic features described in section 3.2 in a leave-one-out cross-validation. The results are compared to the age-only baseline approach for the different resections status in **Table 1**. It can be seen that the proposed supervised PCA approach achieves a higher accuracy and better mean square error than the age-only approach. In contrast, even as the age is included in the feature set, this approach fails on the GTR subset.

4.2. Multivariate Prediction

All methods were cross-validated in a leave-one-out setting, e.g., 100 samples were used to infer the 101th sample for the GTR dataset. From the 101 obtained results, the major test statistics as used in the BraTS challenge were computed: Accuracy (based on the three different time intervals described in section 1), mean squared error (MSE), median error, and Spearman rank correlation. For classifiers, all metrics are computed with respect to the class value (long-survivors: 824 days, mid-survivors: 379 days, short-survivors: 150 days). For the accuracy, we also assessed the statistical significance of the result with a binomial test and provide the p-value. All results can be seen in **Table 2**.

On the GTR subset, no model achieved better results than the age-only baseline. However, on the STR subset, most models were more predictive of survival than the age-only approach.

4.3. Feature Reduction Approaches

Several publications have shown the predictiveness of radiomics for survival prediction on different datasets (see section 1). On top, the re-implemented methods could reveal predictiveness of radiomic features for survival on patients with subtotal resection. However, these methods, as well as different machine learning models presented in this paper and as well as the majority of radiomic approaches submitted in the BraTS challenge 2018 failed on patients that underwent GTR. Thus, in this subsection, we focus on the GTR patients.

Even as all presented models performed worse than the age-only regressor, it can also be observed that the results of different machine models achieve strongly varying results. This could be due to the high number of radiomic features. Thus, it is evaluated whether the two proposed feature reduction techniques can produce more robust outcomes on the GTR dataset.

The presented unsupervised feature subset selection has two subsequent steps: First, the correlation matrix clustering, which suppresses pairwise correlated features, reduced the number of radiomic features from 9,870 to 5,338. Then, the VIF-based feature reduction, that checks also for multicollinearity, further reduced the number of features to 94. Combined with the age,

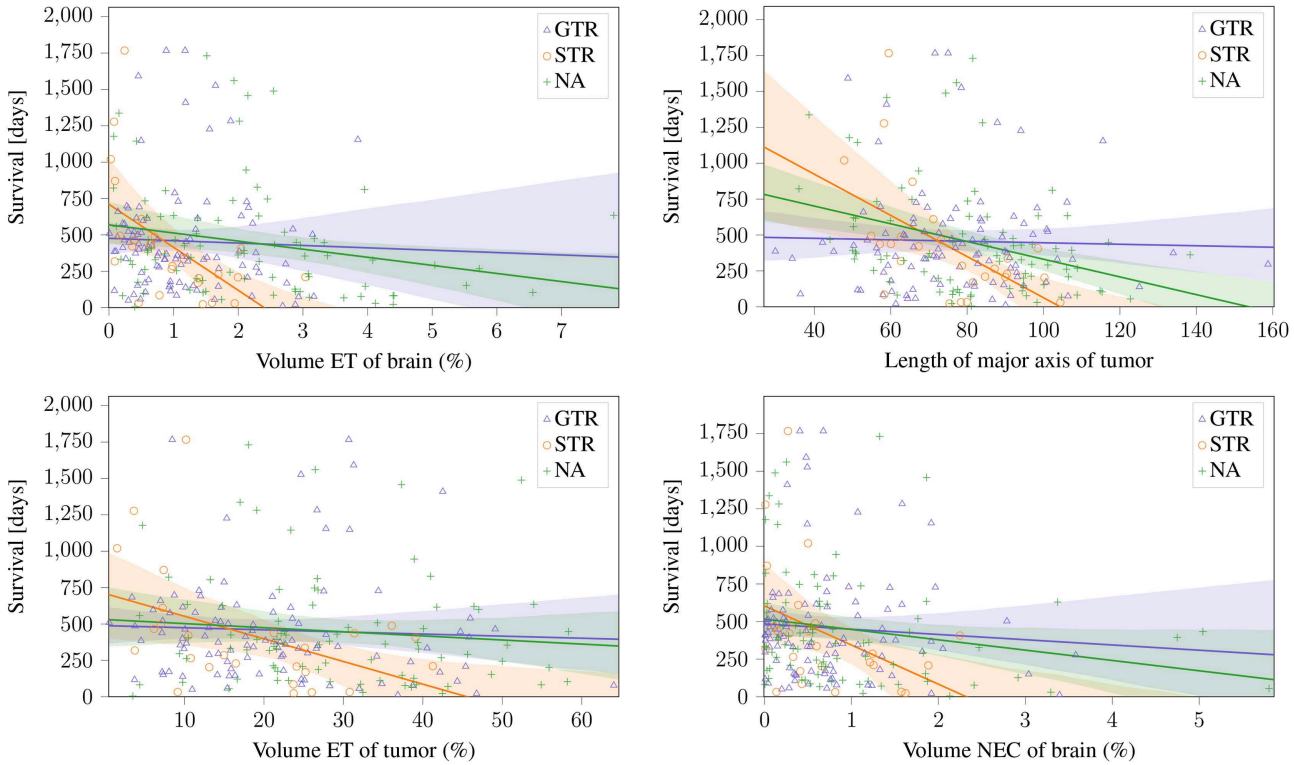


FIGURE 5 | Linear relationship between previously reported significant features and the different resection status.

TABLE 1 | Age-only linear regression compared to the supervised PCA (sPCA) model for the different types of resection status.

Model	Accuracy	p (Binomial)	MSE	Median err.	SpearmanR
GTR					
Age-only	0.48	0.00	109966	159	0.47
sPCA	0.27	0.20	148421	208	−0.41
STR					
Age-only	0.23	0.40	186772	194	−0.64
sPCA	0.46	0.21	148028	205	0.31
NA					
Age-only	0.37	0.49	136866	196	0.29
sPCA	0.26	0.20	159227	231	−0.03

The metrics are explained in section 4.2.

95 features were obtained, which is slightly less than the number of examples.

On this reduced feature set, hypothesis testing is feasible. Without any correction for false positives, 5 of the 95 features would have been considered significant (p <0.05). However, after Benjamini–Hochberg correction, only the age of the patient ($p = 5.8 \times 10^{-5}$) and one radiomic feature, the Wavelet LHH ImageIntensity Kurtosis on the necrotic part in the T2 acquisition remained significant. All statistically significant features can be seen in **Table 3**.

Next, the unsupervised feature selection method based on PCA is considered. After applying PCA as explained in section 3.3, 15 principal components representing tumor shape were kept, 38 for the image intensity statistics features, 55 for the gray level features, and 98 for the wavelet features. The extracted features were concatenated together with the age in order to be used for multivariate prediction.

These features, as well as the features selected by the VIF-analysis, were separately employed for survival prediction (see **Table 4**). Consistent to 4.2, all features were normalized with a robust scaler, subtracting the median and scaling by the by the interquartile range, and the same machine learning models were utilized.

5. DISCUSSION

Previous findings, especially those using volumetric features, could be reproduced for patients with subtotal resection. Furthermore, different considered machine learning models also showed predictiveness of survival. Thus, even as the sample size was limited, and different machine learning models show varying results, radiomic features seem to be correlated to patient survival for patients with subtotal resection.

However, when applying these methods to patients that underwent GTR, no significant relationship between radiomic features and overall survival could be identified. In effect, for this subgroup, the considered previously published and newly

TABLE 2 | Performance comparison of different machine learning models for the different types of resection status.

Model	Accuracy	p (Binomial)	MSE	Median err.	SpearmanR
GTR					
Regression	0.43	0.04	1778965427	394	0.11
Lasso	0.44	0.03	3070271	272	0.24
Ridge	0.42	0.07	1765220107	394	0.11
kNN	0.29	0.40	159451	248	−0.04
RFR	0.36	0.46	154447	190	0.14
SVR	0.27	0.20	140088	189	−0.77
SVC	0.41	0.11	175014	229	0.12
Boruta+RFC	0.45	0.02	133497	229	0.36
STR					
Regression	0.54	0.03	13748742	271	0.32
Lasso	0.46	0.20	415488	228	0.18
Ridge	0.54	0.03	13754344	272	0.32
kNN	0.50	0.09	173798	211	0.21
RFR	0.58	0.01	144744	125	0.43
SVR	0.23	0.40	175720	157	−0.65
SVC	0.31	0.99	221768	229	−0.43
Boruta+RFC	0.50	0.09	92744	229	0.45
NA					
Regression	0.32	0.91	1434873	412	−0.02
Lasso	0.39	0.25	533963	282	0.11
Ridge	0.32	0.91	1435941	412	−0.02
kNN	0.25	0.13	166072	254	−0.18
RFR	0.33	1.00	146459	247	0.20
SVR	0.25	0.13	155827	225	−0.78
SVC	0.40	0.16	169944	229	0.00
Boruta+RF	0.30	0.56	162734	229	−0.13

*All features as described in section 3.5 including the age were used as input. For the age-only approach as comparison, see **Table 1**.*

TABLE 3 | Correlation analysis of VIF—selected features with OS for GTR patients.

Feature	Correlation with OS	*p*-value
Age	−0.46	0.000001
Wavelet LHH ImageIntensity Kurtosis T2 NEC	0.39	0.00002
Wavelet LLL ngtdm Complexity T2 ET	0.22	0.03
Wavelet LHL ImageIntensity Kurtosis T1CE nec	0.22	0.03
Wavelet LLL ImageIntensity Minimum T1 ET	−0.20	0.05

developed radiomic models could not identify any connection between image based features and survival that went beyond the predictiveness of patient age. In previously published findings, the resection status is often not known or not clearly stated (Gutman et al., 2013; Macyszyn et al., 2015; Kickingereder et al., 2016; Lao et al., 2017; Li et al., 2017), or radiomic features are not assessed dependent on resection status (Zhang et al., 2014; Nie et al., 2019). Patient age, a clinical marker that is not strongly predictive of survival for patient without total tumor resection (cf. **Figure 4**) seems to be the strongest predictor of patient survival after GTR. One single feature, the Wavelet LHH

TABLE 4 | Performance comparison of different feature selection methods and machine learning models for GTR patients.

Model	Accuracy	p (Binomial)	MSE	Median err.	SpearmanR
VIF-BASED FEATURE SUBSET					
Regression	0.47	0.01	28154236838	1,112	0.18
Regr. BH	0.46	0.07	109618	148	0.46
Lasso	0.36	0.60	2293591760	557	0.05
Ridge	0.33	1.0	16655918658	672	−0.02
kNN	0.30	0.53	159553	223	−0.08
RFR	0.35	0.75	149299	207	0.15
SVR	0.27	0.20	140181	189	−0.77
SVC	0.40	0.17	194331	445	0.06
FEATURES EXTRACTED by PCA					
Regression	0.39	0.17	672193	478	0.03
Lasso	0.36	0.59	688037	559	0.04
Ridge	0.38	0.25	558488	457	0.02
kNN	0.34	0.92	149014	194	0.07
RFR	0.34	0.92	163826	218	0.02
SVR	0.27	0.20	140298	189	−0.79
SVC	0.40	0.17	198655	445	0.05

Regr. BH, Regression on all features that were significant after Benjamini–Hochberg multiple test correction.

ImageIntensity Kurtosis T2 NEC, was statistically significant after Benjamini–Hochberg correction. However, after leveraging this finding in a predictive regression model, no clear benefit could be observed. Why radiomic features were not predictive on GTR patients remains unclear. It can only be hypothesized that survival for STR patients depends on the malignancy of the primary subtotally resected tumor, while survival for GTR patients relates to possible metastases that are not directly dependent on image features of the original tumor.

It can nevertheless be concluded that OS of brain tumor patients given radiomic images is strongly dependent not only on the preoperative images themselves. Given a high number of features and strong influences that cannot be assessed with preoperative MRI images, survival prediction is an ill-posed problem on a limited dataset. Researchers need to pay attention to the problems that arise when using radiomics or other big data methods on wide data, i.e., datasets with much more features than observations. Specifically, challenge participants and other researchers in clinical data analysis need to be fully aware of overfitting pitfalls, not only on the training set, but even on the validation dataset.

In radiomics, a very high number of features are extracted. In our case, a total of 9,871 features were initially considered. Combined with a limited dataset, as is often the case for medical applications, problems arise due to the curse of dimensionality. One problem encountered is the robustness of significance: The features that are significant on the whole dataset are not necessarily significant on the training subset, and vice versa, features identified as significant on a small dataset do not need to be significant on larger datasets. Although it is impossible to test all possible combinations of different radiomic features and machine learning models, we think that our evaluation shows the

limitations of radiomic analysis on glioblastoma patients with GTR. To be as robust as possible against such subset biases, we used extensive cross-validation and determined an orthogonal subset of features.

Next to the difficulties encountered when applying radiomics to patients that underwent GTR, we believe one main limitation in the BraTS challenge 2018 was the small training dataset in combination with overfitted approaches. In the BraTS setting, a predefined validation set was released by the organizers, and could be used by all contributors to evaluate their algorithms during development. Thus, if contributors test different algorithms or hyperparameter settings on this left-out validation set, it has to be taken into account that one may accidentally overfit on this validation set. Such "result-peeking" invalidates accuracy scores on this left-out dataset, i.e. the developed approaches may generalize poorly to other samples. In fact, it seemed that this actually happened during the BraTS challenge. As can be seen on the official BraTS challenge online leaderboard (Bakas et al., 2018a), a total of nine different teams obtained accuracy scores at least as good as ours on the validation dataset. However, on the test set, our naïve algorithm scored the 3rd place out of 26 participants. On this dataset, segmentation and OS results could not be evaluated by participants, making this part of the data impossible to overfit. In contrast to private datasets, algorithm developers cannot—be it deliberately or accidentally—invalidate the obtained results by result-peeking in such a setting.

Thus, challenges such as the BraTS challenge are important for unbiased algorithm comparison and to assess whether findings from research are robust and can be applied to translational medicine. Here, it was assessed whether findings in radiomics of glioblastoma patients can be transferred to patients that underwent GTR. In this case, classical radiomic features seem not to be suited for robust results in survival prediction. In contrast, positive findings, with previously reported approaches as well as with different machine learning techniques can be reported for patients with subtotal resection.

Nevertheless, the approaches presented in this paper are not exhaustive. We do not want to present the new "best" survival prediction algorithm. Default parameter settings were utilized for all machine learning techniques, as exhaustive hyperparameter tuning—as employed by most winning approaches in machine learning challenges—on a small dataset would invalidate the results. The approaches presented in this manuscript, especially those relying on orthogonal feature subset selection, were utilized to analyze the robustness of radiomic features. They may not be the "best" algorithms for survival prediction. Thus, C-index, hazard ratio, or KM analysis were not regarded, as the focus of this analysis lies on robustness of radiomic features, and not on a single survival prediction algorithm.

6. CONCLUSION

The BraTS survival prediction challenge focuses on glioblastoma patients that underwent GTR. This paper shows that adding information from radiomic features to the age of the patient does not necessarily improve accuracy for this task. To show this, we evaluated different published techniques as well as a sophisticated radiomic feature extraction combined with modern machine learning techniques. However, no helpful information could be extracted, and our baseline—a linear regression on the age of the patient—could not be consistently outperformed on this limited dataset. In contrast, on patients with a different resection status—either where the resection status was not available or the tumor was subtotally resected—previously published findings could be reproduced, and different machine learning techniques could extract information predictive for overall survival.

In order to move from fundamental research to translational medicine, future research in brain tumor radiomics should focus on finding novel radiomic features that are applicable if the patient undergoes surgery. A possible set of features that was not assessed in this study are location based features. Location based features are not as established as shape or texture features in radiomics. However, they could be more promising for survival prediction even for patients that underwent GTR, as the position of the tumor in the brain could influence prognosis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2018/registration.html.

ETHICS STATEMENT

This work relies only on the BraTS dataset. For the use of this dataset, no ethics statement is necessary.

AUTHOR CONTRIBUTIONS

LW performed algorithm development and implementation and wrote the manuscript. CH assisted in algorithm conception and interpretation of the results. DM supervised the work and critically revised the manuscript.

FUNDING

This work was supported by the International Research Training Group 2150 and the German Research Foundation (DFG) under grant no ME3737/3-1.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00073/full#supplementary-material

REFERENCES

Baid, U., Talbar, S., Rane, S., Gupta, S., Thakur, M. H., Moiyadi, A., et al. (2019). "Deep learning radiomics algorithm for gliomas (drag) model: a novel approach using 3d unet based deep convolutional neural network for predicting survival in gliomas," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 369–379.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG Collection*. Technical report. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM Collection*. Technical report. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Min Ha, S., and Bergman, M. (2018a). *BraTS Validation Survival Leaderboard 2018*. Available online at: https://web.archive.org/web/20191019144601/https://www.cbica.upenn.edu/BraTS18//lboardValidationSurvival.html

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018b). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv [Pre-print]. arXiv:1811.02629*. Available online at: http://arxiv.org/abs/1811.02629

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Front Neurosci.* 11:543. doi: 10.3389/fnins.2017.00543

Cho, H.-H., Lee, S.-H., Kim, J., and Park, H. (2018). Classification of the glioma grading using radiomics analysis. *PeerJ.* 6:e5982. doi: 10.7717/peerj.5982

Feng, X., Tustison, N., and Meyer, C. (2019). "Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features" in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 279–288.

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: images are more than pictures, they are data. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Gutman, D. A., Cooper, L. A. D., Hwang, S. N., Holder, C. A., Gao, J., Aurora, T. D., et al. (2013). MRI imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* 267, 560–569. doi: 10.1148/radiol.13120118

Jain, R., Poisson, L. M., Gutman, D., Scarpace, L., Hwang, S. N., Holder, C. A., et al. (2014). Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology* 272, 484–493. doi: 10.1148/radiol.14131691

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer Publishing Company, Incorporated.

Jones, E., Oliphant, T., and Peterson, P. (2001). *SciPy: Open Source Scientific Tools for Python*.

Jungo, A., McKinley, R., Meier, R., Knecht, U., Vera, L., Perez-Beteta, J., et al. (2018). "Towards uncertainty-assisted brain tumor segmentation and survival prediction," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 474–485.

Kickingereder, P., Burth, S., Wick, A., Götz, M., Eidel, O., Schlemmer, H.-P., et al. (2016). Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology* 280, 880–889. doi: 10.1148/radiol.2016160845

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J. Stat. Softw.* 36, 1–13. doi: 10.18637/jss.v036.i11

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J., et al. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci. Rep.* 7:10353. doi: 10.1038/s41598-017-1649-8

Li, Q., Bai, H., Chen, Y., Sun, Q., Liu, L., Zhou, S., et al. (2017). A fully-automatic multiparametric radiomics model: towards reproducible and prognostic imaging signature for prediction of overall survival in glioblastoma multiforme. *Sci. Rep.* 7:14331. doi: 10.1038/s41598-017-14753-7

Macyszyn, L., Akbari, H., Pisapia, J. M., Da, X., Attiah, M., Pigrish, V., et al. (2015). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro Oncol.* 18, 417–425. doi: 10.1093/neuonc/nov127

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Nie, D., Lu, J., Zhang, H., Adeli, E., Wang, J., Yu, Z., et al. (2019). Multi-channel 3d deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Sci. Rep.* 9:1103. doi: 10.1038/s41598-018-37387-9

Ostrom, Q. T., Gittleman, H., Truitt, G., Boscia, A., Kruchko, C., and Barnholtz-Sloan, J. S. (2018). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2011–2015. *Neuro Oncol.* 20(Suppl 4):iv1–iv86. doi: 10.1093/neuonc/noy131

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Puybareau, E., Tochon, G., Chazalon, J., and Fabrizio, J. (2019). "Segmentation of gliomas and prediction of patient overall survival: a simple and fast procedure," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 199–209.

Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., et al. (2018). Radiomics: the facts and the challenges of image analysis. *Eur. Radiol. Exp.* 2:36. doi: 10.1186/s41747-018-0068-z

Shboul, Z. A., Vidyaratne, L., Alam, M., and Iftekharuddin, K. M. (2018). "Glioblastoma and survival prediction," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 358–368.

Sun, L., Zhang, S., and Luo, L. (2019). "Tumor segmentation and survival prediction in glioma with deep learning," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 83–93.

Suter, Y., Jungo, A., Rebsamen, M., Knecht, U., Herrmann, E., Wiest, R., et al. (2019). "Deep learning versus classical regression for brain tumor patient survival prediction," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 429–440.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4itk: improved n3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Voorhees, E. M. (1986). *Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval*. Technical report, Ithaca, NY.

Weninger, L., Rippel, O., Koppers, S., and Merhof, D. (2019). "Segmentation of brain tumors and patient survival prediction: Methods for the brats 2018 challenge," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 3–12.

Zacharaki, E., Morita, N., Bhatt, P., O'Rourke, D., Melhem, E., and Davatzikos, C. (2012). Survival analysis of patients with high-grade gliomas based on data mining of imaging variables. *Am. J. Neuroradiol.* 33, 1065–1071. doi: 10.3174/ajnr.A2939

Zhang, Z., Jiang, H., Chen, X., Bai, J., Cui, Y., Ren, X., et al. (2014). Identifying the survival subtypes of glioblastoma by quantitative volumetric analysis of MRI. *J. Neuro Oncol.* 119, 207–214. doi: 10.1007/s11060-014-1478-2

Zhou, M., Scott, J., Chaudhury, B., Hall, L., Goldgof, D., Yeom, K., et al. (2018). Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *Am. J. Neuroradiol.* 39, 208–216. doi: 10.3174/ajnr.A5391

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

REVIEW
published: 11 December 2019
doi: 10.3389/fncom.2019.00083

Data Augmentation for Brain-Tumor Segmentation: A Review

Jakub Nalepa[1,2]*, *Michal Marcinkiewicz*[3] and *Michal Kawulok*[2]

[1] Future Processing, Gliwice, Poland, [2] Silesian University of Technology, Gliwice, Poland, [3] Netguru, Poznan, Poland

Data augmentation is a popular technique which helps improve generalization capabilities of deep neural networks, and can be perceived as implicit regularization. It plays a pivotal role in scenarios in which the amount of high-quality ground-truth data is limited, and acquiring new examples is costly and time-consuming. This is a very common problem in medical image analysis, especially tumor delineation. In this paper, we review the current advances in data-augmentation techniques applied to magnetic resonance images of brain tumors. To better understand the practical aspects of such algorithms, we investigate the papers submitted to the Multimodal Brain Tumor Segmentation Challenge (BraTS 2018 edition), as the BraTS dataset became a standard benchmark for validating existent and emerging brain-tumor detection and segmentation techniques. We verify which data augmentation approaches were exploited and what was their impact on the abilities of underlying supervised learners. Finally, we highlight the most promising research directions to follow in order to synthesize high-quality artificial brain-tumor examples which can boost the generalization abilities of deep models.

Keywords: MRI, image segmentation, data augmentation, deep learning, deep neural network

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Rong Pan,
Arizona State University, United States
Guotai Wang,
University of Electronic Science and Technology of China, China

***Correspondence:**
Jakub Nalepa
jnalepa@ieee.org

Received: *30 April 2019*
Accepted: *27 November 2019*
Published: *11 December 2019*

Citation:
Nalepa J, Marcinkiewicz M and Kawulok M (2019) Data Augmentation for Brain-Tumor Segmentation: A Review.
Front. Comput. Neurosci. 13:83.
doi: 10.3389/fncom.2019.00083

1. INTRODUCTION

Deep learning has established the state of the art in many sub-areas of computer vision and pattern recognition (Krizhevsky et al., 2017), including medical imaging and medical image analysis (Litjens et al., 2017). Such techniques automatically discover the underlying data representation to build high-quality models. Although it is possible to utilize generic priors and exploit domain-specific knowledge to help improve representations, deep features can capture very discriminative characteristics and explanatory factors of the data which could have been omitted and/or unknown for human practitioners during the process of manual feature engineering (Bengio et al., 2013).

In order to successfully build well-generalizing deep models, we need huge amount of ground-truth data to avoid overfitting of such large-capacity learners, and "memorizing" training sets (LeCun et al., 2016). It has become a significant obstacle which makes deep neural networks quite challenging to apply in the medical image analysis field where acquiring *high-quality* ground-truth data is time-consuming, expensive, and very human-dependent, especially in the context of brain-tumor delineation from magnetic resonance imaging (MRI) (Isin et al., 2016; Angulakshmi and Lakshmi Priya, 2017; Marcinkiewicz et al., 2018; Zhao et al., 2019). Additionally, the majority of manually-annotated image sets are imbalanced—examples belonging to some specific classes are often under-represented. To combat the problem of limited medical training sets, data augmentation techniques, which generate synthetic training examples, are being actively developed in the literature (Hussain et al., 2017; Gibson et al., 2018; Park et al., 2019).

In this review paper, we analyze the brain-tumor segmentation approaches available in the literature, and thoroughly investigate which techniques have been utilized by the participants of the Multimodal Brain Tumor Segmentation Challenge (BraTS 2018). To the best of our knowledge, the dataset used for the BraTS challenge is currently the largest and the most comprehensive brain-tumor dataset utilized for validating existent and emerging algorithms for detecting and segmenting brain tumors. Also, it is heterogeneous in the sense that it includes both low- and high-grade lesions, and the included MRI scans have been acquired at different institutions (using different MR scanners). We discuss the brain-tumor data augmentation techniques already available in the literature, and divide them into several groups depending on their underlying concepts (section 2). Such MRI data augmentation approaches have been applied to augment other datasets as well, also acquired for different organs (Amit et al., 2017; Nguyen et al., 2019; Oksuz et al., 2019).

In the BraTS challenge, the participants are given multi-modal MRI data of brain-tumor patients (as already mentioned, both low- and high-grade gliomas), alongside the corresponding ground-truth multi-class segmentation (section 3). In this dataset, different sequences are co-registered to the same anatomical template and interpolated to the same resolution of 1 mm^3. The task is to build a supervised learner which is able to generalize well over the unseen data which is released during the testing phase. In section 4, we summarize the augmentation methods reported in 20 papers published in the BraTS 2018 proceedings. Here, we focused on those papers which *explicitly* mentioned that the data augmentation had been utilized, and clearly stated what kind of data augmentation had been applied. Although such augmentations are single-modal—meaning that they operate over the MRI from a single sequence—they can be easily applied to co-registered series, hence to augment multi-modal tumor examples. Finally, the paper is concluded in section 5, where we summarize the advantages and disadvantages of the reviewed augmentation techniques, and highlight the promising research directions which emerge from (not only) BraTS.

2. DATA AUGMENTATION FOR BRAIN-TUMOR SEGMENTATION

Data augmentation algorithms for brain-tumor segmentation from MRI can be divided into the following main categories (which we render in a taxonomy presented in **Figure 1**): the algorithms exploiting various transformations of the original data, including *affine* image transformations (section 2.1), *elastic* transformations (section 2.2), *pixel-level* transformations (section 2.3), and various approaches for *generating artificial data* (section 2.4). In the following subsections, we review the approaches belonging to all groups of such augmentation methods in more detail.

Traditionally, data augmentation approaches have been applied to increase the size of training sets, in order to allow large-capacity learners benefit from more representative training data (Wong et al., 2016). There is, however, a new trend in the deep learning literature, in which examples are augmented on the fly (i.e., during the inference), in the *test-time*[1] augmentation process. In **Figure 2**, we present a flowchart in which both training- and test-time data augmentation is shown. Test-time data augmentation can help increase the *robustness* of a trained model by simulating the creation of a homogeneous ensemble, where $(n + 1)$ models (of the same type, and trained over the same training data) *vote* for the final class label of an incoming test example, and n denotes the number of artificially-generated samples, elaborated for the test example which is being classified. The robustness of a deep model is often defined as its ability to correctly classify previously unseen examples—such incoming examples are commonly "noisy" or slightly "perturbed" when confronted with the original data, therefore they are more challenging to classify and/or segment (Rozsa et al., 2016). Test-time data augmentation can be exploited for estimating the level of uncertainty of deep networks during the inference—it brings new exciting possibilities in the context of medical image analysis, where quantifying the robustness and deep-network reliability are crucial practical issues (Wang et al., 2019). This type of data augmentation can utilize those methods which *modify* an incoming example, e.g., by applying affine, pixel-level or elastic transformations in the case of brain-tumor segmentation from MRI.

2.1. Data Augmentation Using Affine Image Transformations

In the *affine approaches*, existent image data undergo different operations (rotation, zooming, cropping, flipping, or translations) to increase the number of training examples (Pereira et al., 2016; Liu et al., 2017). Shin et al. pointed out that such traditional data augmentation techniques fundamentally produce very correlated images (Shin et al., 2018), therefore can offer very little improvements for the deep-network training process and future generalization over the unseen test data (such examples do not regularize the problem sufficiently). Additionally, they can also generate anatomically incorrect examples, e.g., using rotation. Nevertheless, affine image transformations are trivial to implement (in both 2D and 3D), they are fairly flexible (due to their hyper-parameters), and are widely applied in the literature. In an example presented in **Figure 3**, we can see that applying simple data augmentation techniques can lead to a significant increase in the number of training samples.

2.1.1. Flip and Rotation

Random flipping creates a mirror reflection of an original image along one (or more) selected axis. Usually, natural images can be flipped along the horizontal axis, which is not the case for the vertical one because up and down parts of an image are not always "interchangeable." A similar property holds for MRI brain images—in the axial plane a brain has two hemispheres, and the brain (in most cases) can be considered anatomically symmetrical. Flipping along the horizontal axis swaps the left

[1]Test-time augmentation is also referred to as the *inference-time* and the *online* data augmentation in the literature.

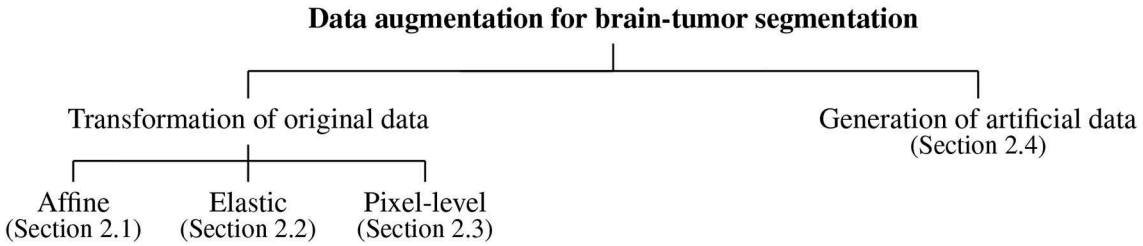


FIGURE 1 | Data augmentation for brain-tumor segmentation—a taxonomy.

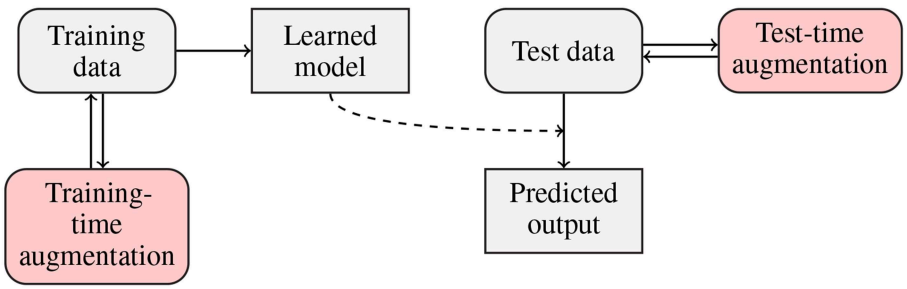


FIGURE 2 | Flowchart presenting training- and test-time data augmentation. In the training-time data augmentation approach, we generate synthetic data to increase the representativeness of a training set (and ultimately build better models), whereas in test-time augmentation, we benefit from the ensemble-like technique, in which multiple homogeneous classifiers vote for the final class label for an incoming example by classifying this sample and a number of its augmented versions.

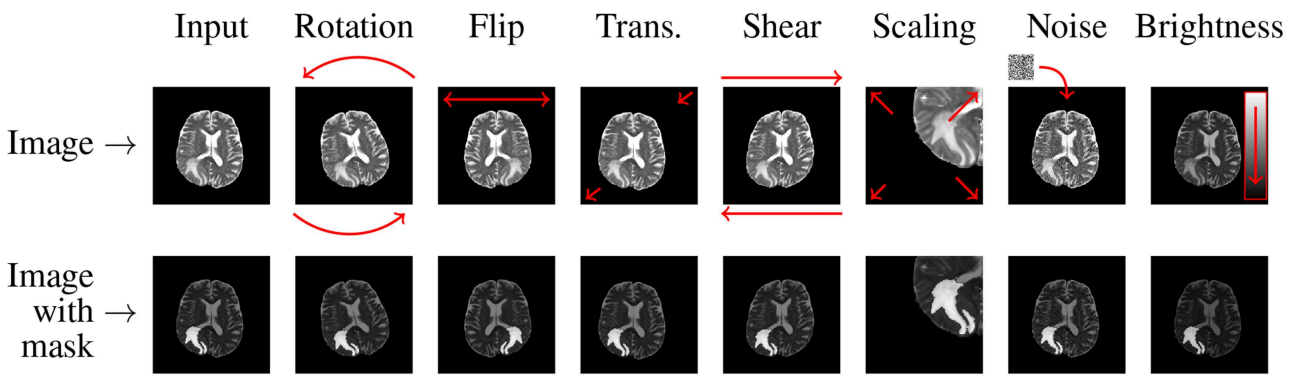


FIGURE 3 | Applying affine and pixel-level (discussed in more detail in section 2.3) transformations can help significantly increase the size (and potentially representativeness) of training sets. In this example, we generate seven new images based on the original MRI (coupled with its ground truth in the bottom row).

hemisphere with the right one, and vice versa. This operation can help various deep classifiers, especially those benefitting from the contextual tumor information, be invariant with respect to their position within the brain which would be otherwise difficult for not representative training sets (e.g., containing brain tumors located only in the left or right hemisphere). Similarly, rotating an image by an angle α around the center pixel can be exploited in this context. This operation is followed by appropriate interpolation to fit the original image size. The rotation operation denoted as R is often coupled with zero-padding applied to the missing pixels:

$$R = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}. \quad (1)$$

2.1.2. Translation

The translation operation shifts the entire image by a given number of pixels in a chosen direction, while applying padding accordingly. It allows the network to not become focused on features present mainly in one particular spatial region, but it forces the model to learn spatially-invariant features instead. As in the case of rotation—since the MRI scans of different patients available in training sets are often not co-registered—translation of an image by a given number of pixels along a selected axis (or axes) can create useful and viable images. However, this procedure may not be "useful" for all deep architectures—convolutional neural networks exploit convolutions and pooling operations, which are intrinsically spatially-invariant (Asif et al., 2018).

2.1.3. Scaling and Cropping

Introducing scaled versions of the original images into the training set can help the deep network learn valuable deep features independently of their original scale. This operation S can be performed independently in different directions (for brevity, we have only two dimensions here):

$$S = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}, \tag{2}$$

and the scaling factors are given as s_x and s_y for the x and y directions, respectively. As tumors vary in size, scaling can indeed bring viable augmented images into a training set. Since various deep architectures require images of the constant size, scaling is commonly paired with cropping to maintain the original image dimensions. Such augmented brain-tumor examples may manifest tumoral features at different scales. Also, cropping can limit the field of view only to those parts of the image which are important (Menze et al., 2015).

2.1.4. Shearing

The shear transformation (H) displaces each point in an image in a selected direction. This displacement is proportional to its distance from the line which goes through the origin and is parallel to this direction:

$$H = \begin{pmatrix} 1 & h_x \\ h_y & 1 \end{pmatrix}, \tag{3}$$

where h_x and h_y denote the shear coefficient in the x and y directions, respectively (as previously, we consider two dimensions for readability). Although this operation can deform shapes, it is rarely used to augment medical image data because we often want to preserve original shape characteristics (Frid-Adar et al., 2018).

2.2. Data Augmentation Using Elastic Image Transformations

Data augmentation algorithms based on unconstrained *elastic transformations* of training examples can introduce shape variations. They can bring lots of noise and damage into the training set if the deformation field is seriously varied—see an example by Mok and Chung (2018) in which a widely-used elastic transform produced a totally unrealistic synthetic MRI scan of a human brain. If the simulated tumors were placed in "unrealistic" positions, it would likely force the segmentation engine to become invariant to contextual information and rather focus on the lesion's appearance features (Dvornik et al., 2018). Although there are works which indicate that such aggressive augmentation may deteriorate the performance of the models in brain-tumor delineation (Lorenzo et al., 2019), it is still an open issue. Chaitanya et al. (2019) showed that visually non-realistic synthetic examples can improve the segmentation of cardiac MRI and noted that it is slightly counter-intuitive—it may have occurred due to the inherent structural and deformation-related characteristics of the cardiovascular system. Finally, elastic transformations often benefit from B-splines (Huang and Cohen, 1996; Gu et al., 2014) or random deformations (Castro et al., 2018).

Diffeomophic mappings play an important role in brain imaging, as they are able to preserve topology and generate biologically plausible deformations. In such transformations, the *diffeomorphism* ϕ (also referred to as a *diffeomorphic mapping*) is given in the spatial domain Ω of a source image I, and transforms I to the target image J: $I \circ \phi^{-1}(\mathbf{x}, 1)$. The mapping is the solution of the differential equation:

$$\frac{d\phi(\mathbf{x}, t)}{dt} = v\left(\phi(\mathbf{x}, t), t\right), \tag{4}$$

where $\phi(\mathbf{x}, 0) = \mathbf{x}$, v is a time-dependent smooth velocity field, $v : \Omega \times t \rightarrow \mathcal{R}^d$, $\phi(\mathbf{x}, t)$ is a geodesic path (d denotes the dimensionality of the spatial domain Ω), and $\phi(\mathbf{x}, t) : \Omega \times t \rightarrow \Omega$. In Nalepa et al. (2019a), we exploited the directly manipulated free-form deformation, in which the velocity vector fields are regularized using B-splines (Tustison et al., 2009). The d-dimensional update field $\delta v_{i_1,\ldots,i_d}$ is

$$\delta v_{i_1,\ldots,i_d} = \frac{\sum_{c=1}^{N_\Omega} \left(\frac{\partial \xi}{\partial \mathbf{x}}\right)_c \prod_{j=1}^{d} B_{i_j}(x_j^c) \prod_{j=1}^{d} B_{i_j}^2(x_j^c)}{\left(\sum_{c=1}^{N_\Omega} \prod_{j=1}^{d} B_{i_j}^2(x_j^c)\right)\left(\sum_{k_1=1}^{r+1} \cdots \sum_{k_d=1}^{r+1} \prod_{j=1}^{d} B_{k_j}^2(x_j^c)\right)}, \tag{5}$$

and $B(\cdot)$ are the B-spline basis functions, N_Ω denotes the number of pixels in the domain of the reference image, r is the spline order (in all dimensions), and $\frac{\partial \xi}{\partial \mathbf{x}}$ is the gradient of the spatial similarity metric at a pixel c. The B-spline functions act as regularizers of the solution for each parametric dimension (Tustison and Avants, 2013).

Examples of brain-tumor images generated using diffeomorphic registration are given in **Figure 4**—such artificially-generated data significantly improved the abilities of deep learners, especially when combined with affine transformations, as we showed in Nalepa et al. (2019a). The generated (I') images preserve topological information of the original image data (I) with subtle changes to the tissue. Diffeomorphic registration may be applied not only to images exposing anatomical structures (Tward and Miller, 2017). In **Figure 5**, we present examples of simple shapes which underwent this transformation—the topological information is clearly maintained in the generated images as well.

2.3. Data Augmentation Using Pixel-Level Image Transformations

There exist augmentation techniques which do not alter geometrical shape of an image (therefore, all geometrical features remain unchanged during the augmentation process), but affect the pixel intensity values (either locally, or across the entire image). Such operations can be especially useful in medical image analysis, where different training images are acquired in different locations and using different scanners, hence can be intrinsically heterogeneous in the pixel intensities, intensity

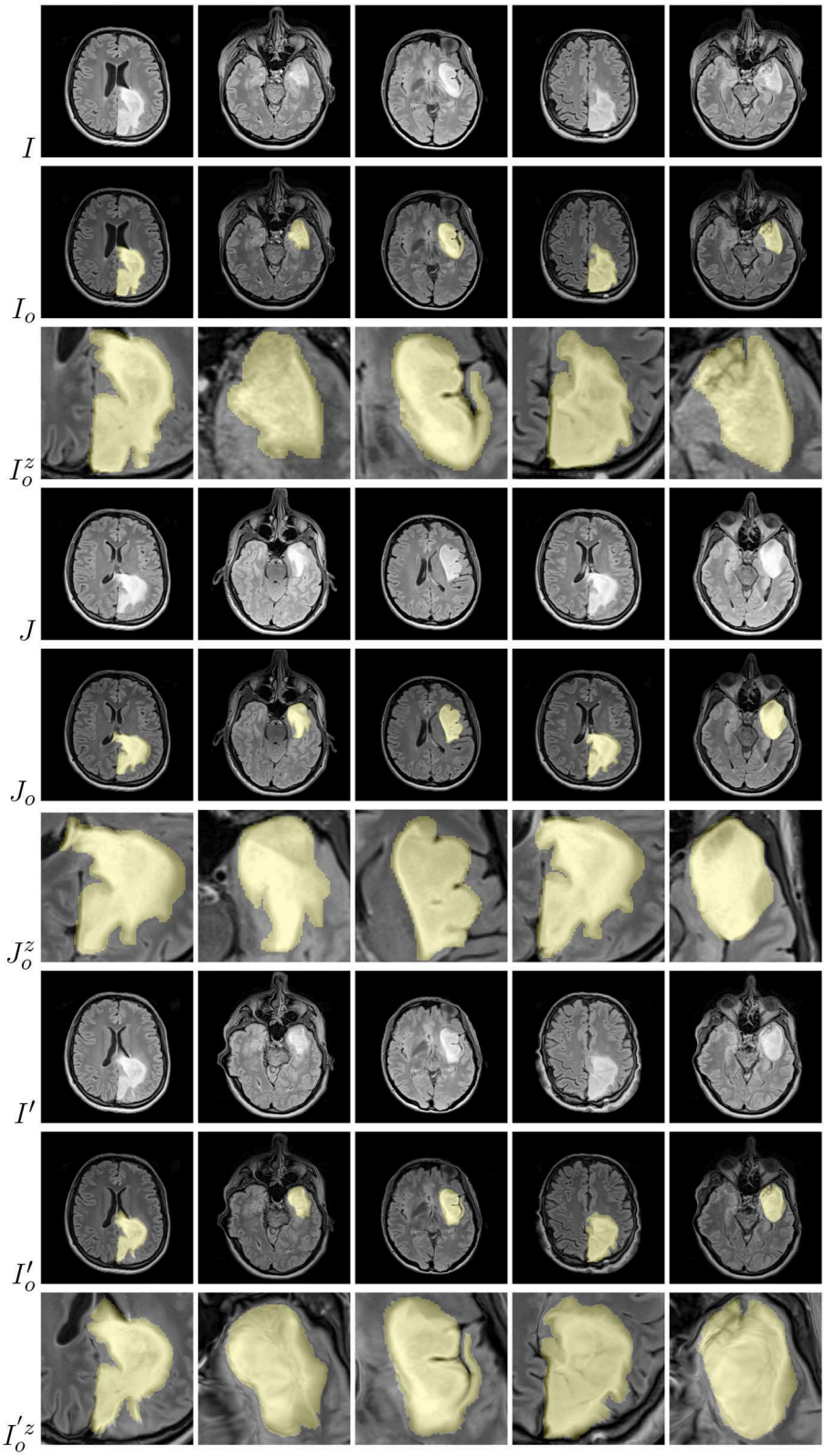


FIGURE 4 | Diffeomorphic image registration applied to example brain images allowed for obtaining visually-plausible generated images. For source (*I*), target (*J*), and artificially generated (*I′*) images, we also present tumor masks overlayed over the corresponding original images (in yellow; rows with the *o* subscript), alongside a zoomed part of a tumor (rows with the *z* superscript).

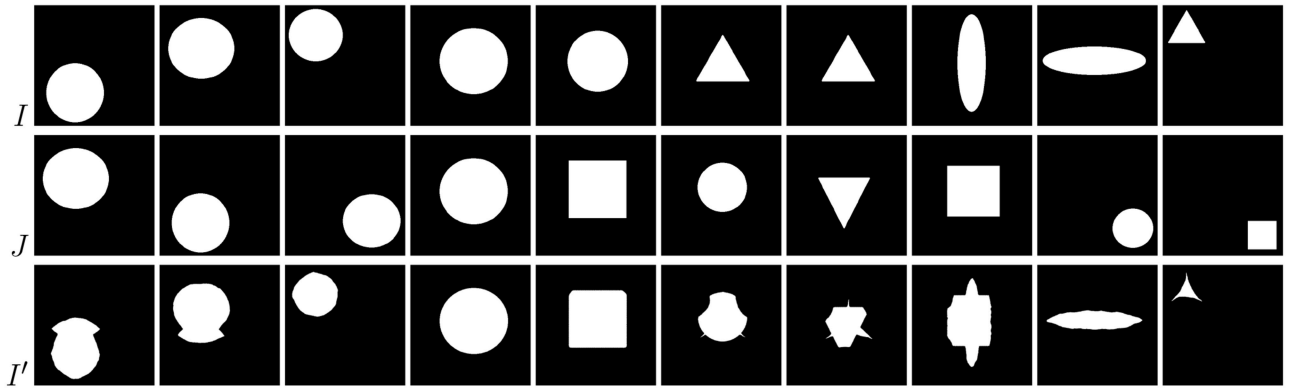


FIGURE 5 | Diffeomorphic image registration applied to basic shapes which underwent simple affine registration (translation) before diffeomorphic mapping. Source images (*I*) transformed to match the corresponding targets (*J*) still clearly expose their spatial characteristics (*I′*).

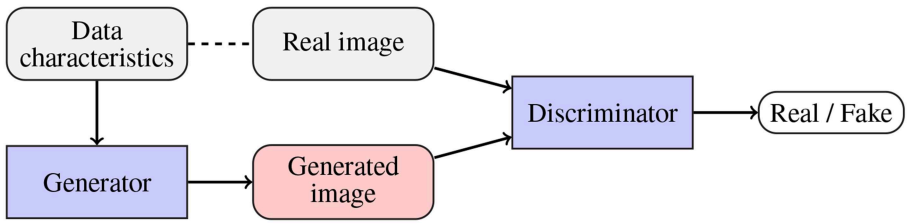


FIGURE 6 | Generative adversarial networks are aimed at generating fake data (by a *generator*; potentially using some available data characteristics) which is indistinguishable from the original data by the *discriminator*. Therefore, the generator and discriminator compete with one another.

gradients or "saturation"[2]. During the *pixel-level augmentation*, the pixel intensities are commonly perturbed using either *random* or zero-mean *Gaussian* noise (with the standard deviation corresponding to the appropriate data dimension), with a given probability (the former operation is referred to as the *random intensity variation*). Other pixel-level operations include shifting and scaling of pixel-intensity values (and modifying the image brightness), applying gamma correction and its multiple variants (Agarwal and Mahajan, 2017; Sahnoun et al., 2018), sharpening, blurring, and more (Galdran et al., 2017). This kind of data augmentation is often exploited for high-dimensional data, as it can be conveniently applied to selected dimensions (Nalepa et al., 2019b).

2.4. Data Augmentation by Generating Artificial Data

To alleviate the problems related to the basic data augmentation approaches (including the problem of generating correlated data samples), various approaches toward *generating artificial data* (GAD) have been proposed. Generative adversarial networks (GANs), originally introduced in Goodfellow et al. (2014), are being exploited to augment medical datasets (Han et al., 2019; Shorten and Khoshgoftaar, 2019). The main objective of a GAN (**Figure 6**) is to generate a new data example (by a *generator*) which will be indistinguishable from the real data by the *discriminator* (the generator competes with the discriminator, and the overall optimization mimics the min-max game). Mok and Chung proposed a new GAN architecture which utilizes a coarse-to-fine generator whose aim is to capture the manifold of the training data and generate augmented examples (Mok and Chung, 2018). Adversarial networks have been also used for semantic segmentation of brain tumors (Rezaei et al., 2017), brain-tumor detection (Varghese et al., 2017), and image synthesis of different modalities (Yu et al., 2018). Although GANs allow us to introduce invariance and robustness of deep models with respect to not only affine transforms (e.g., rotation, scaling, or flipping) but also to some shape and appearance variations, convergence of the adversarial training and existence of its equilibrium point remain the open issues. Finally, there exist scenarios in which the generator renders multiple very similar examples which cannot improve the generalization of the system—it is known as the *mode collapse problem* (Wang et al., 2017).

An interesting approach for generating phantom image data was exploited in Gholami et al. (2018), where the authors utilized a multi-species partial differential equations (PDE) growth model of a tumor to generate synthetic lesions. However, such data does not necessarily follow the correct intensity distribution of a real MRI, hence it should be treated as a separate modality, because using the artificial data which is sampled from a very different distribution may adversely affect the overall segmentation performance by "tricking" the underlying deep

[2]These variations can be however alleviated by appropriate data standardization.

model (Wei et al., 2018). The tumoral growth model itself captured the time evolution of enhancing and necrotic tumor concentrations together with the edema induced by a tumor. Additionally, the deformation of a lesion was simulated by incorporating the linear elasticity equations into the model. To deal with the different data distributions, the authors applied CycleGAN (Zhu et al., 2017) for performing domain adaptation (from the generated phantom data to the real BraTS MRI scans). The experimental results showed that the domain adaptation was able to generate images which were practically indistinguishable from the real data, therefore could be safely included in the training set.

A promising approach of combining training samples using their linear combinations (referred to as *mixup*) was proposed by Zhang et al. (2017), and further enhanced for medical image segmentation by Eaton-Rosen et al. in their *mixmatch* algorithm (Eaton-Rosen et al., 2019), which additionally introduced a technique of selecting training samples that undergo linear combination. Since the medical image datasets are often imbalanced (with the tumorous examples constituting the minority class), training patches with highest "foreground amounts" (i.e., the number of pixels annotated as tumorous) are combined with those with the lowest concentration of foreground. The authors showed that their approach can increase performance in medical-image segmentation tasks, and related its success to the mini-batch training. It is especially relevant in the medical-image analysis, because the sizes of input scans are usually large, hence the batches are small to keep the training memory requirements feasible in practice. Such data-driven augmentation techniques can also benefit from growing ground-truth datasets (e.g., BraTS) which manifest large variability of brain tumors, to generate even more synthetic examples. Also, they could be potentially applied at test time to build an ensemble-like model, if a training patch/image which matches the test image being classified was efficiently selected from the training set.

3. DATA

In this work, we analyzed the approaches which were exploited by the BraTS 2018 participants to segment brain tumors from MRI (45 methods have been published, Crimi et al., 2019), and verified which augmentation scenarios were exploited in these algorithms. All of those techniques have been trained over the BraTS 2018 dataset consisting of MRI-DCE data of 285 patients with diagnosed gliomas: 210 patients with high-grade glioblastomas (HGG), and 75 patients with low-grade gliomas (LGG), and validated using the validation set of 66 previously unseen patients (both LGG and HGG, however the grade has not been revealed) (Menze et al., 2015; Bakas et al., 2017a,b,c). Each study was manually annotated by one to four expert readers. The data comes in four co-registered modalities: native pre-contrast (T1), post-contrast T1-weighted (T1c), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). All the pixels have one of four labels attached: healthy tissue, Gd-enhancing tumor (ET), peritumoral edema (ED), the necrotic and non-enhancing tumor core (NCR/NET). The scans were skull-stripped and interpolated to the same shape (155, 240, 240 with the voxel size of 1 mm^3).

Importantly, this dataset manifests very heterogeneous image quality, as the studies were acquired across different institutions, and using different scanners. On the other hand, the delineation procedure was clearly defined which allowed for obtaining similar ground-truth annotations across various readers. To this end, the BraTS dataset—as the largest, most heterogeneous, and carefully annotated set—has been established as a standard brain-tumor dataset for quantifying the performance of existent and emerging detection and segmentation approaches. This heterogeneity is pivotal, as it captures a wide range of tumor characteristics, and the models trained over BraTS are easily applicable for segmenting other MRI scans (Nalepa et al., 2019).

To show this desirable feature of the BraTS set experimentally, we trained our U-Net-based ensemble architecture (Marcinkiewicz et al., 2018) using (a) BraTS 2019 training set (exclusively FLAIR sequences) and (b) our set of 41 LGG (WHO II) brain-tumor patients who underwent the MR imaging with a MAGNETOM Prisma 3T system (Siemens, Erlangen, Germany) equipped with a maximum field gradient strength of 80 mT/m, and using a 20-channel quadrature head coil. The MRI sequences were acquired in the axial plane with a field of view of 230 $\times$ 190 mm, matrix size 256 $\times$ 256 and 1 mm slice thickness with no slice gap. In particular, we exploited exclusively FLAIR series with TE = 386 ms, TR = 5,000 ms, and inversion time of 1,800 ms for segmentation of brain tumors. These scans underwent the same pre-processing as applied in the case of BraTS, however they were *not* segmented following the same delineation protocol, hence the characteristics of the manual segmentation likely differ across (a) and (b). The 4-fold cross-validation showed that although the deep models trained over (a) and (b) gave the statistically different results at $p < 0.001$, according to the two-tailed Wilcoxon test[3], the ensemble of models trained over (a) correctly detected 71.4% (5/7 cases) of brain tumors in the WHO II test dataset, which included seven patients kept aside while building an ensemble, with the average whole-tumor DICE of 0.80, where DICE is given as:

$$\mathrm{DICE(A, B)} = \frac{2 \cdot |A \cap B|}{|A| + |B|}, \tag{6}$$

where A and B are two segmentations, i.e., manual and automated, $0 \leq \mathrm{DICE} \leq 1$, and $\mathrm{DICE} = 1$ means the perfect segmentation score. On the other hand, a deep model trained over the WHO II training set and used for segmenting the test WHO II cases detected 85.7% tumors (6/7 patients) with the average whole-tumor DICE = 0.84. This tiny experiment shows that the segmentation engines trained over BraTS can capture tumor characteristics which are manifested in MRI data acquired and analyzed using different protocols, and allow us to obtain high-quality segmentation. Interestingly, if we train our ensemble over the combined BraTS 2019 and WHO II training sets, we

[3]We tested the null hypothesis saying that applying the models trained exclusively over the BraTS or our WHO II datasets leads to the same-quality segmentation.

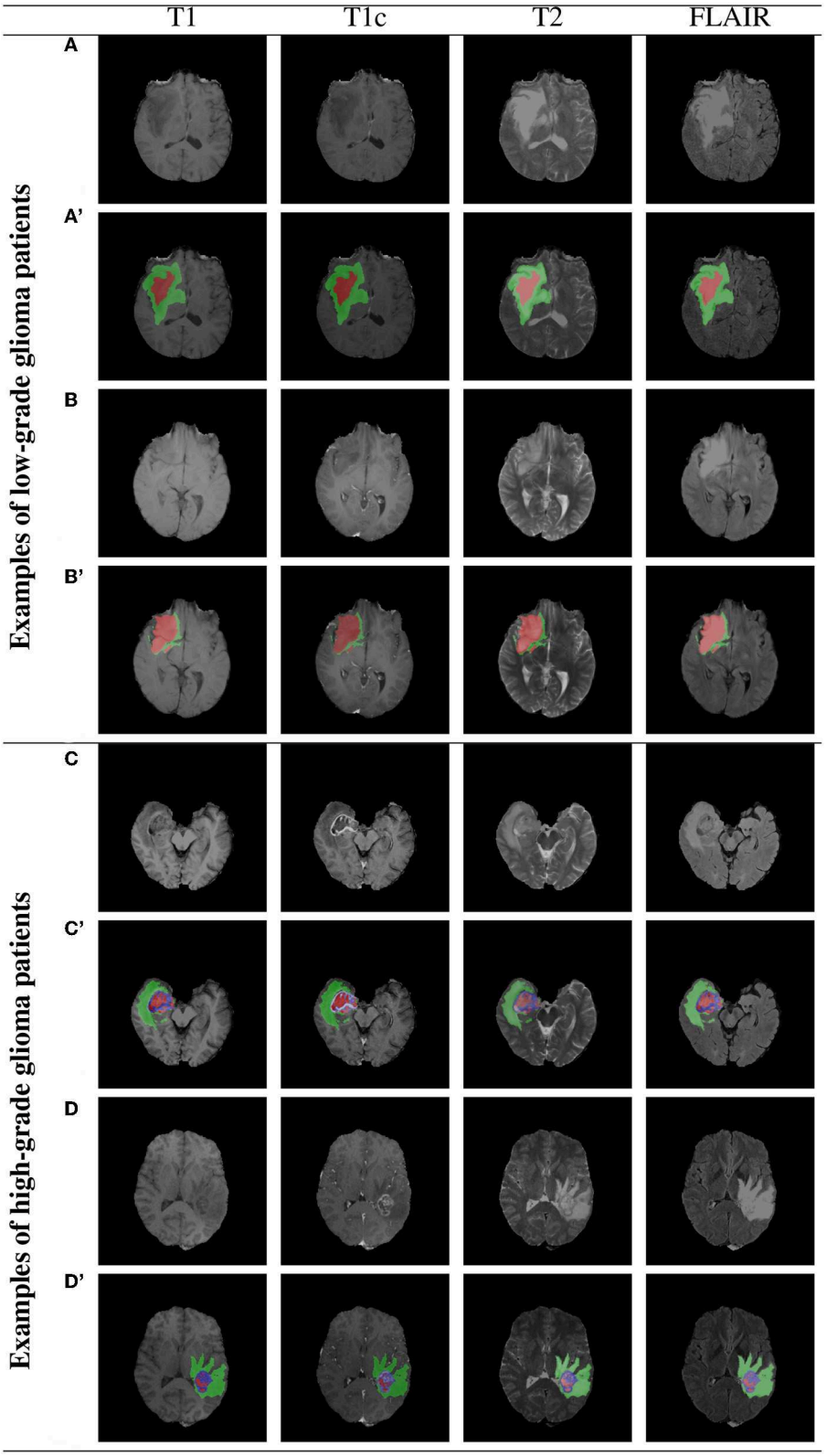


FIGURE 7 | Two example low- and high-grade glioma patients from the BraTS 2018 dataset: red—GD-enhancing tumor (ET), green—peritumoral edema (ED), and blue—necrotic and non-enhancing tumor core (NCR/NET); **(A–D)** show original images, whereas **(A'–D')** present overlaid ground-truth masks.

TABLE 1 | Data augmentation techniques applied in the approaches validated within the BraTS 2018 challenge framework.

References	Model	Flip	Rot.	Trans.	Scale	Shear	Elastic	GAD	Pixel-wise
Albiol et al., 2019	VGG, Inception, Dense	3D affine transformations							
Benson et al., 2018*	CNN (encoder-decoder)	Yes							Random
Carver et al., 2018	U-Net	Yes							
Chandra et al., 2018	V-Net, ResNet-18, FC-CRF	Yes			Yes				
Dai et al., 2018	Domain-adapted U-Net	Yes							
Feng et al., 2018	U-Net	Yes							
Gholami et al., 2018*	U-Net							PDE	
Isensee et al., 2018	U-Net	Yes	Yes		Yes		Random		Gamma
Kao et al., 2018	DeepMedic, 3D U-Net	Yes							
Kermi et al., 2018	U-Net	Yes	Yes	Yes					
Lachinov et al., 2018*	Cascaded U-Net	Yes					B-spline		Gaussian
Ma and Yang, 2018	3D CNN	Yes	Yes		Yes				
McKinley et al., 2018	Dense CNN	Yes	Yes						Shift, scale
Mehta and Arbel, 2018	U-Net		Yes	Yes	Yes	Yes			
Myronenko, 2018	CNN (encoder-decoder)	Yes			Yes				Shift
Nuechterlein and Mehta, 2018	3D-ESPNet	Yes			Yes				
Puybareau et al., 2018	VGG-16		Yes		Yes				
Rezaei et al., 2018†	Voxel-GAN		Yes		Yes				Gaussian
Sun et al., 2018	CNN, DFKZ, 3D CNN	Yes							Gaussian
Wang et al., 2018*†	CNN	Yes	Yes		Yes				Random
Number of methods utilizing this augmentation→15			8	2	9	1	2	1	8
Percentage (%) of methods utilizing this augmentation→75			40	10	45	5	10	5	40

The top-performing techniques (over the unseen test set) are annotated with green.

**The authors verified the impact of data augmentation of the generalization abilities of their deep models.*

†*The authors used both training- and test-time data augmentation.*

will end up having the correct detection of 85.7% tumors (6/7 cases) with the average whole-tumor DICE of 0.76. We can appreciate the fact that we were able to improve the detection, but the segmentation quality slightly dropped, showing that the detected case was challenging to segment. Finally, it is worth mentioning that this experiment sheds only some light on the effectiveness of applying the deep models (or other data-driven techniques) trained over BraTS for analyzing different MRI brain images. The manual delineation protocols were different, and the lack of inter-rater agreement may play pivotal role in quantifying automated segmentation algorithms over such differently acquired and analyzed image sets—it is unclear if the differences result from the inter-rater disagreement of the incorrect segmentation (Hollingworth et al., 2006; Fyllingen et al., 2016; Visser et al., 2019).

3.1. Example BraTS Images

Example BraTS 2018 images are rendered in **Figure 7** (two low-grade and two high-grade glioma patients), alongside the corresponding multi-class ground-truth annotations. We can appreciate that different parts of the tumors are manifested in different modalities—e.g., necrotic and non-enhancing tumor core is typically hypo-intense in T1-Gd when compared to T1 (Bakas et al., 2018). Therefore, multi-modal analysis appears crucial to fully benefit from the available image information.

4. BRAIN-TUMOR DATA AUGMENTATION IN PRACTICE

4.1. BraTS 2018 Challenge

The BraTS challenge is aimed at evaluating the state-of-the-art approaches toward accurate multi-class brain-tumor segmentation from MRI. In this work, we review all *published* methods which were evaluated within the framework of the BraTS 2018 challenge—although 61 teams participated in the testing phase (Bakas et al., 2018), only 45 methods were finally described and published in the post-conference proceedings (Crimi et al., 2019). We verify which augmentation techniques were exploited to help boost generalization abilities of the proposed supervised learners. We exclusively focus on 20 papers (44% of all manuscripts) in which the authors *explicitly* stated that the augmentation had been used and report the type of the applied augmentation.

In **Table 1**, we summarize all investigated brain-tumor segmentation algorithms, and report the deep models utilized in the corresponding works alongside the augmentation techniques. In most of the cases, the authors followed the cross-validation scenario, and divided the training set into multiple non-overlapping folds. Then, separate models were trained over such folds, and the authors finally formed an ensemble of heterogeneous classifiers (trained over different training data) to segment previously unseen test brain-tumor images. Also, there

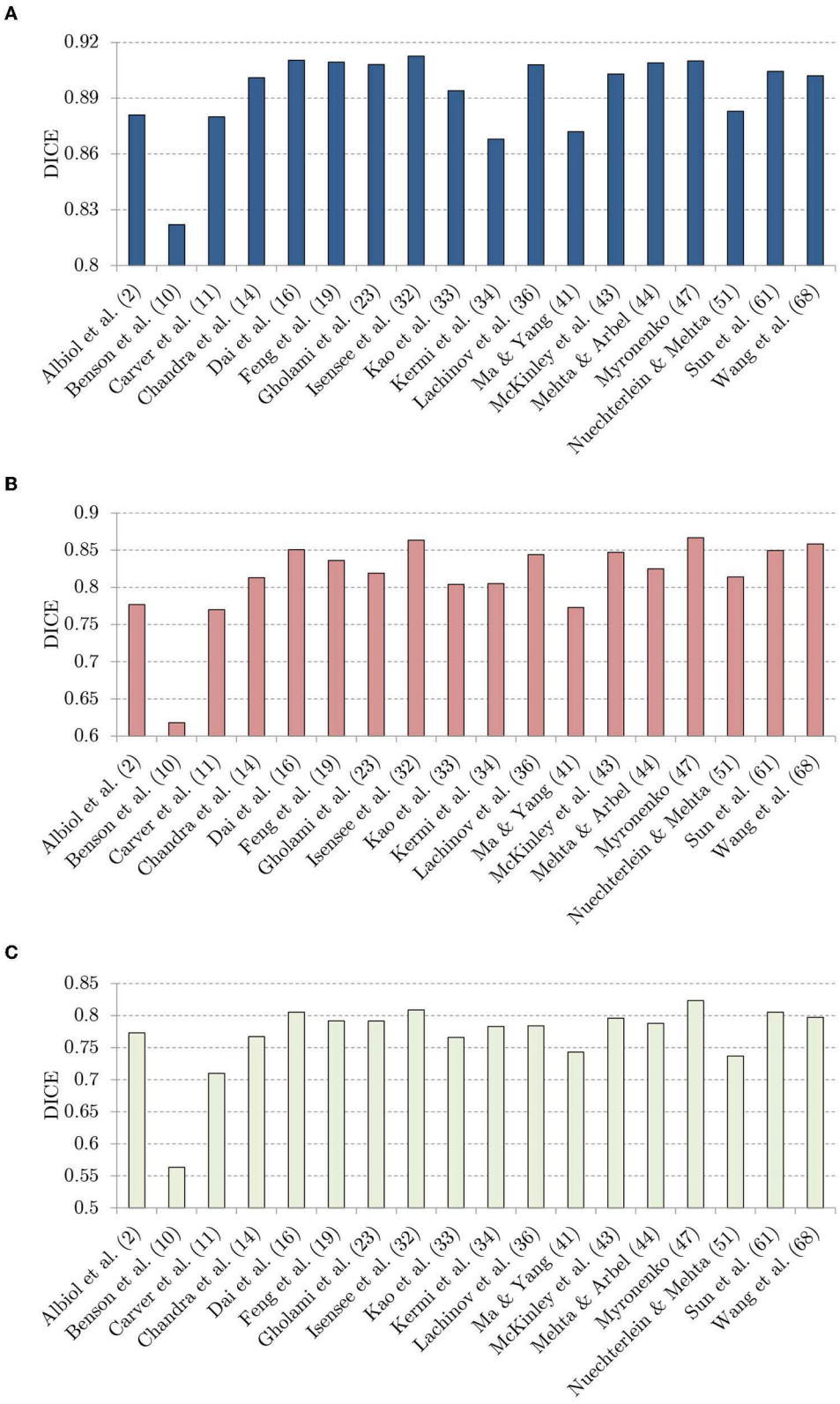


FIGURE 8 | The DICE values: **(A)** whote-tumor (WT), **(B)** tumor core (TC), and **(C)** enhancing tumor (ET), obtained using the investigated techniques over the BraTS 2018 validation set.

TABLE 2 | The impact of applying data augmentation on the average DICE scores.

References	Without augmentation			With augmentation			Change (in %)		
	WT	TC	ET	WT	TC	ET	ΔWT	ΔTC	ΔET
Benson et al., 2018	0.82	0.64	0.59	0.82	0.61	0.56	0	−5	−5
Gholami et al., 2018	0.89	0.80	0.74	0.91	0.82	0.79	+2	+3	+7
Lachinov et al., 2018*	0.91	0.84	0.77	0.91	0.84	0.78	0	0	+1
Wang et al., 2018	0.90	0.85	0.79	0.90	0.86	0.80	0	+1	+1

For the methods reported by Lachinov et al. (2018) and Wang et al. (2018), we analyzed the best-performing models.
**The authors verified the impact of data augmentation over the training set.*

TABLE 3 | The impact of applying data augmentation on the average Hausdorff distance values (in mm).

References	Without augmentation			With augmentation			Change (in %)		
	WT	TC	ET	WT	TC	ET	ΔWT	ΔTC	ΔET
Benson et al., 2018	94.28	130.70	18.12	13.57	17.95	14.29	−86	−86	−21
Wang et al., 2018	5.38	6.61	3.34	6.18	6.37	3.13	+26	−4	−6

For the method reported by Wang et al. (2018), we analyzed the best-performing models. Note that Gholami et al. (2018) and Lachinov et al. (2018) did not present the Hausdorff distances obtained using their approaches.

TABLE 4 | The fully convolutional neural networks proposed in Lorenzo et al. (2019) have been trained using a number of datasets with different preprocessing and augmentations.

Setup→	A, A'	B, B'	C, C'	D, D'	E, E'	F, F'	G, G'	H, H'	I, I'	J, J'	K, K'	L, L'	M, M'	N, N'	O, O'
Feature centering	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Vertical flip	No	No	Yes	No	Yes	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Horizontal flip	No	No	No	Yes	Yes	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Max. rotation ($\measuredangle_{max}$)	0	0	0	0	0	45	45	45	45	45	90	90	90	90	90
Augmentation factor	1, 2	1, 2	2, 4	2, 4	4, 8	2, 4	2, 4	4, 8	4, 8	8, 16	2, 4	2, 4	4, 8	4, 8	8, 16

In the prime versions, we applied elastic deformations. This table comes from our previous paper (Lorenzo et al., 2019).

are approaches, e.g., by Albiol et al. (2019), Chandra et al. (2018), or Sun et al. (2018), in which a variety of deep neural architectures were used.

In the majority of investigated brain-tumor segmentation techniques, the authors applied relatively simple training-time data augmentation strategies—the combination of training- and test-time augmentation was used only in two methods (Rezaei et al., 2018; Wang et al., 2018). In 75% of the analyzed approaches, random flipping was executed to increase the training set size and provide anatomically correct brain images[4]. Similarly, rotating and scaling MRI images was applied in 40% and 45% of techniques, respectively. Since modern deep network architectures are commonly translation-invariant, this type of affine augmentation was used only in two works. Although other augmentation strategies were not as popular as easy-to-implement affine transformations, it is worth noting that the pixel-wise operations were utilized in all of the top-performing techniques (the algorithms by Myronenko (2018), Isensee et al. (2018), and McKinley et al. (2018) achieved the first, second, and third place across all segmentation algorithms[5], respectively). Additionally, Isensee et al. (2018) exploited elastic transformations in their aggressive data augmentation procedure which significantly increased the size and representativeness of their training sets, and ultimately allowed for outperforming a number of other learners. Interestingly, the authors showed that the state-of-the-art U-Net architecture can be extremely competitive with other (much deeper and complex) models if the data is appropriately curated. It, in turn, manifests the importance of data representativeness and quality in the context of robust medical image analysis.

In **Figure 8**, we visualize the DICE scores obtained using almost all investigated methods (Puybareau et al., 2018; Rezaei et al., 2018 did not report the results over the unseen BraTS 2018 validation set, therefore these methods are not included in the figure). It is worth mentioning that the trend is fairly coherent for all classes (whole tumor, tumor core, and enhancing tumor), and the best-performing methods by Isensee et al. (2018), McKinley

[4]Note that we do not count the algorithm proposed by Albiol et al. (2019), because the authors were not very specific about their augmentation strategies.

[5]For more detail on the validation and scoring procedures, see Bakas et al. (2018).

et al. (2018), and Myronenko (2018) consistently outperform the other techniques in all cases. Although the success of these approaches obviously lies not only in the applied augmentation techniques, it is notable that the authors extensively benefit from generating additional synthetic data.

Albeit data augmentation is introduced in order to improve the generalization capabilities of supervised learners, this impact was verified only in four BraTS 2018 papers (Benson et al., 2018; Gholami et al., 2018; Lachinov et al., 2018; Wang et al., 2018). Gholami et al. (2018) showed that their PDE-based augmentation delivers very significant improvement in the DICE scores obtained for segmenting all parts of the tumors in the multi-class classification. The same performance boost (in the DICE values obtained for each class) was reported by Lachinov et al. (2018). Finally, Wang et al. (2018) showed that the proposed test-time data augmentation led to improving the performance of their convolutional neural networks.

In **Table 2**, we gathered the DICE scores obtained with and without the corresponding data augmentation, alongside the change in DICE (reported in %; the larger the DICE score becomes, the better segmentation has been obtained). Interestingly, training-time data augmentation appeared to be adversely affecting the performance of the algorithm presented by Benson et al. (2018). On the other hand, the authors showed that the Hausdorff distance, being the maximum distance of

TABLE 5 | Five best-performing configurations of our fully convolutional neural network according to the Friedman's test (at $p < 0.05$) taking into account the results elaborated for the WHO II validation set (Lorenzo et al., 2019).

Variant→	I	E	O	E'	J'
Rank	4.75	5.50	6.00	7.25	7.75

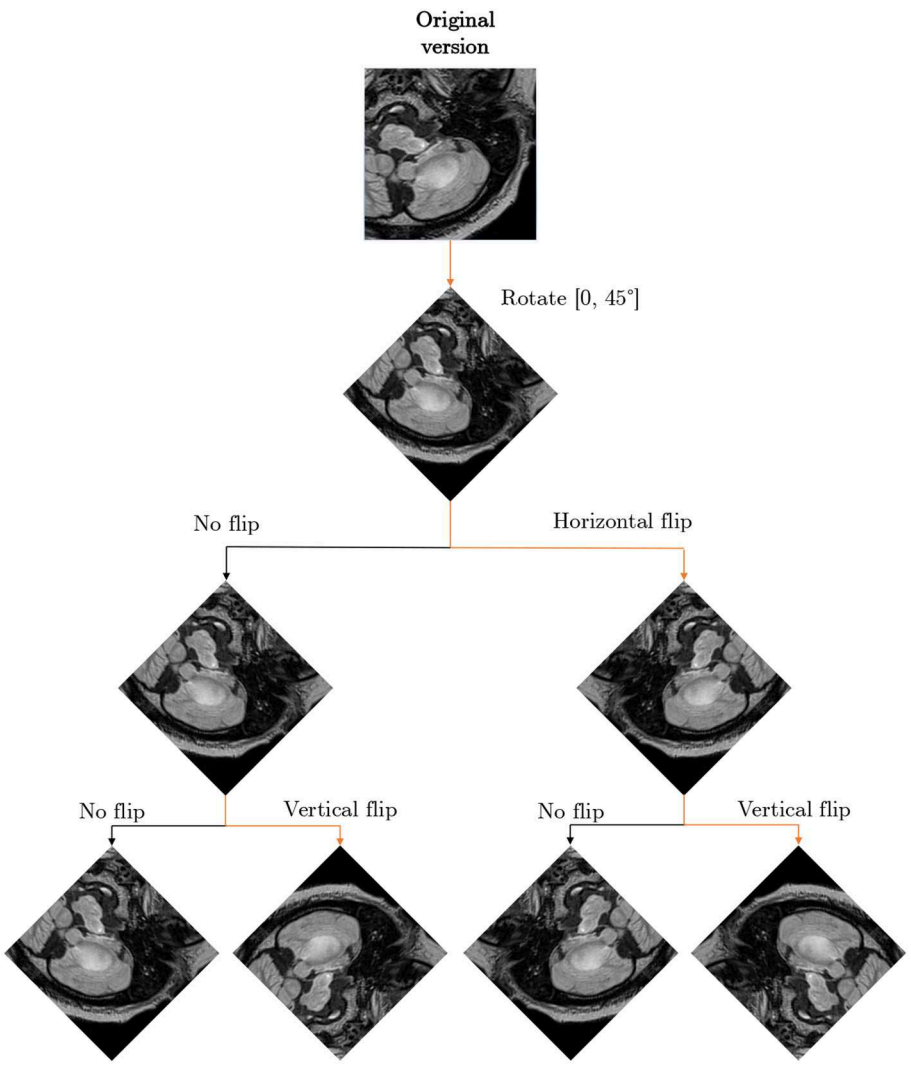


FIGURE 9 | Exploiting various augmentations and coupling them into an augmentation tree allow us to generate multiple versions of an original patch (or image) which may be included in a training set. This figure is inspired by Lorenzo et al. (2019).

all points from the segmented lesion to the corresponding nearest point of the ground-truth segmentation (Sauwen et al., 2017), significantly dropped, hence the maximum segmentation error quantified by this metric was notably reduced (the smaller the Hausdorff distance becomes, the better segmentation has been elaborated; **Table 3**). Test-time data augmentation exploited by Wang et al. (2018) not only decreased DICE for the whole-tumor segmentation, but also caused the increase of the correspoding Hausdorff distance. Therefore, applying it in the WT segmentation scenario led to decreasing the abilities of the underlying models. Overall, the vast majority of methods neither report nor analyze the real impact of the incorporated augmentation techniques on the classification performance and/or inference time of their deep models. Although we believe the authors did investigate the advantages (and disadvantages) of their data generation strategies (either experimentally or theoretically), data augmentation is often used a standard tool which is applied to any *difficult* data (e.g., imbalanced, with highly under-represented classes).

TABLE 6 | The results, both (a) average, and (b) median DICE over our clinical MRI data of low-grade glioma (WHO II) patients in the whole-tumor segmentation task, for different augmentation scenarios.

	Augmentation	Training	Validation	Test
(a)	Without	0.823	0.743	0.763
	Flip	0.836	0.790	0.785
	DIR	0.858	0.777	0.773
	DIR + Flip	**0.865**	**0.808**	**0.800**
(b)	Without	0.823	0.779	0.785
	Flip	0.838	0.808	0.797
	DIR	0.859	0.802	0.792
	DIR + Flip	**0.867**	**0.816**	**0.809**

The results come from our paper (Nalepa et al., 2019a). The best results are boldfaced.

4.2. Beyond the BraTS Challenge

Although practically all brain-tumor segmentation algorithms which emerge in the recent literature have been tested over the BraTS datasets, we equipped our U-Nets with a battery of augmentation techniques (summarized in **Table 4**) and verified their impact over our clinical MRI data in Lorenzo et al. (2019). In this experiment, we have focused on the whole-tumor segmentation, as it was an intermediate step in the automated dynamic contrast-enhanced MRI analysis, in which perfusion parameters have been extracted for the entire tumor volume. Additionally, this dataset was manually delineated by a reader (8 years of experience) who highlighted the whole-tumor areas only.

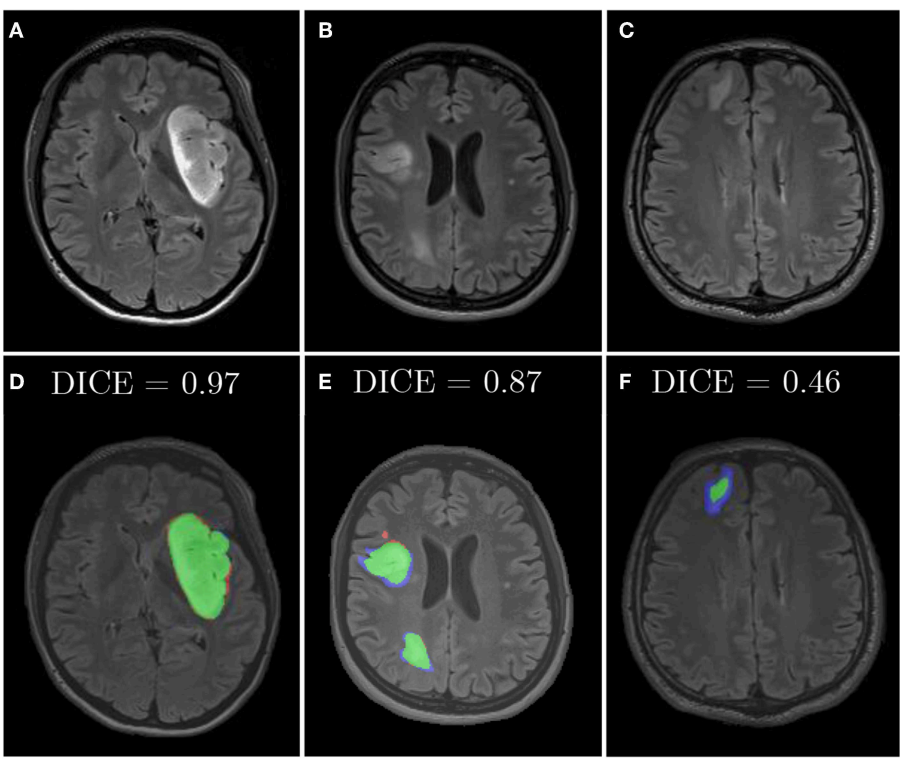


FIGURE 10 | Examples from our clinical dataset segmented using our deep network trained in the DIR+Flip setting: **(A–C)** are original images, **(D–F)** are corresponding segmentations. Green color represents true positives, blue—false negatives, and red—false positives.

We executed multi-step augmentation by applying both affine and elastic deformations of tumor examples, and increased the cardinality of our training sets up to 16×. In **Figure 9**, we can observe how executing simple affine transformations leads to new synthetic image patches. Since various augmentation approaches may be utilized at different depths of this augmentation tree, the number of artificial examples can be significantly increased. The multi-fold cross-validation experiments showed that introducing rotated training examples was pivotal to boost the generalization abilities of underlying deep models. To verify the statistical importance of the results, we executed the Friedman's ranking tests which revealed that the horizontal flip with additional rotation is crucial to build well-generalizing deep learners in the patch-based segmentation scenario (**Table 5**).

Similarly, we applied diffeomorphic image registration (DIR) coupled with a recommendation algorithm[6] to select training image pairs for registration in the data augmentation process (Nalepa et al., 2019a). The proposed augmentation was compared with random horizontal flipping, and the experiments indicated that the combined approach leads to statistically significant (Wilcoxon test at $p < 0.01$) improvements in DICE (**Table 6**). In **Figure 10**, we have gathered example segmentations obtained using our DIR+Flip deep model, alongside the corresponding DICE values. Although the original network, trained over the original training set would correctly detect and segment large tumors (**Figures 10A,B**), it failed for relatively small lesions which were under-represented in the training set (**Figure 10C**). Similarly, synthesizing artificial training examples helped improving the performance of our models in the case of brain tumors located in the brain areas which have not been originally included in the dataset (by applying rotation and flipping).

5. CONCLUSION

In this paper, we reviewed the state-of-the-art data augmentation methods applied in the context of segmenting brain tumors from MRI. We carefully investigated all BraTS 2018 papers and analyzed data augmentation techniques utilized in these methods. Our investigation revealed that the affine transformations are still the most widely-used in practice, since they are trivial to implement and can elaborate anatomically-correct brain-tumor examples. There are, however, augmentation methods which combine various approaches, also including elastic transformations. A very interesting research direction encompasses algorithms which can generate artificial images (e.g., based on the tumoral growth models) that not necessarily follow real-life data distribution, but can be followed by other techniques to ensure correctness of such phantoms. The results showed that data augmentation was pivotal in the best-performing BraTS algorithms, and Isensee et al. (2018) experimentally proved that well-known and widely-used fully-convolutional neural networks can outperform other (perhaps much more deeper and complex) learners, if the training data is appropriately cleansed and curated. It clearly indicates the importance of introducing effective data augmentation methods for medical image data, which benefit from affine transformations (in 2D and 3D), pixel-wise modifications and elastic transform to deal with the problem of limited ground-truth data. In **Table 7**, we gather the advantages and disadvantages of all groups of brain-tumor data augmentation techniques analyzed in this review. Finally, these approaches can be easily applied in both single- and multi-modal scans,

TABLE 7 | The pros and cons of state-of-the-art brain-tumor data augmentation algorithms.

Advantages	Disadvantages
Transformation of original data	
Affine transformations	
• Easy to implement and understand	• Produce correlated images
• Operate in real-time due to low time complexity	• Easily generate anatomically incorrect examples (*)
• Applicable in training- and test-time	
• Deliver invariance with respect to the lesion position, scale, and rotation	
Elastic transformations	
• Can be applicable in training- and test-time	• Not trivial to implement
• Can introduce variations in shape	• Often have high time complexity
	• Easily generate anatomically incorrect examples (*)
Pixel-wise transformations	
• Easy to implement and understand	• Cannot introduce changes in shape
• Operate in real-time due to low time complexity	
• Applicable in training- and test-time	
• Can simulate different acquisition scenarios	
Generation of artificial data	
• Can synthesize realistic examples	• (Very) high time complexity
• (Potentially) applicable in test-time	• GANs applicable in training-time only
• Can introduce invariance with respect to affine transformations and appearance variations	• Can easily render multiple similar examples (mode collapse problem)

**The real impact of incorporating unrealistic examples into training sets still needs investigation.*

[6]We used a recommendation algorithm for selecting source-target image pairs that undergo registration. Such pairs should contain the training images which capture lesions positioned in the same or close part of the brain, as the totally different images can easily render unrealistic brain-tumor examples. A potential drawback of this recommendation technique is its time complexity which amounts to $\mathcal{O}(||\mathbf{T}||^2)$, where $||\mathbf{T}||$ is the cardinality of the original training set.

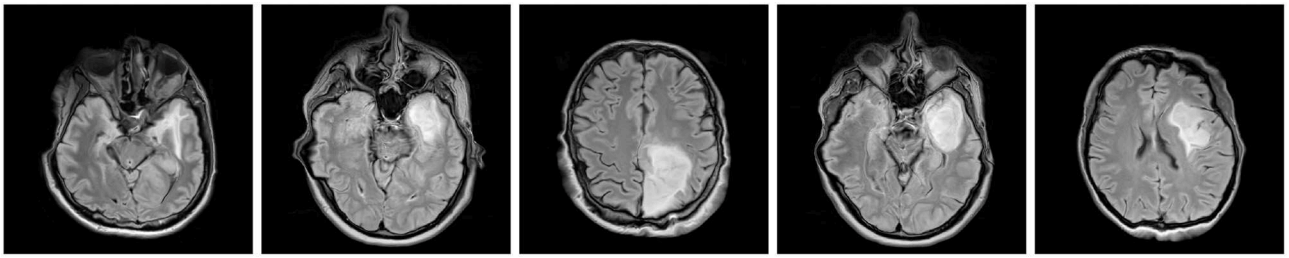


FIGURE 11 | Anatomically incorrect brain images may still manifest valid tumor features—the impact of including such examples (which may be easily rendered by various data-generation augmentation techniques) into training sets for brain-tumor detection and segmentation tasks is yet to be revealed.

usually by synthesizing artificial examples separately for each image modality.

Although data augmentation became a pivotal part of virtually all deep learning-powered methods for segmenting brain lesions (due to the lack of very large, sufficiently heterogeneous and representative ground-truth sets, with BraTS being an exception), there are still promising and unexplored research pathways in the literature. We believe that hybridizing techniques from various algorithmic groups, introducing more data-driven augmentations, and applying them at training- and test-time can further boost the performance of large-capacity learners. Also, investigating the impact of including not necessarily anatomically correct brain-tumor scans into training sets remains an open issue (see the examples of anatomically incorrect brain images which still manifest valid tumor characteristics in **Figure 11**).

AUTHOR CONTRIBUTIONS

JN designed the study, performed the experiments, analyzed data, and wrote the manuscript. MM provided selected implementations and experimental results, and contributed to writing of some parts of the initial version of the manuscript. MK provided qualitative segmentation analysis and visualizations.

FUNDING

This work was supported by the Polish National Centre for Research and Development under the Innomed Grant (POIR.01.02.00-00-0030/15). JN was supported by the Silesian University of Technology funds (The Rector's Habilitation Grant No. 02/020/RGH19/0185). The research undertaken in this project led to developing Sens.AI—a tool for automated segmentation of brain lesions from T2-FLAIR sequences (https://sensai.eu). MK was supported by the Silesian University of Technology funds (Grant No. 02/020/BK_18/0128).

ACKNOWLEDGMENTS

The authors are grateful to the Reviewers for their constructive and valuable comments that helped improve the paper. JN thanks Dana K. Mitchell for lots of inspiring discussions on (not only) brain MRI analysis.

This paper is in memory of Dr. Grzegorz Nalepa, an extraordinary scientist, pediatric hematologist/oncologist, and a compassionate champion for kids at Riley Hospital for Children, Indianapolis, USA, who helped countless patients and their families through some of the most challenging moments of their lives.

REFERENCES

Agarwal, M., and Mahajan, R. (2017). Medical images contrast enhancement using quad weighted histogram equalization with adaptive gama correction and homomorphic filtering. *Proc. Comput. Sci.* 115, 509–517. doi: 10.1016/j.procs.2017.09.107

Albiol, A., Albiol, A., and Albiol, F. (2019). "Extending 2D deep learning architectures to 3D image segmentation problems," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 73–82.

Alex, V., Mohammed Safwan, K. P., Chennamsetty, S. S., and Krishnamurthi, G. (2017). "Generative adversarial networks for brain lesion detection," in *Medical Imaging 2017: Image Processing*, eds M. A. Styner and E. D. Angelini (SPIE), 113–121. doi: 10.1117/12.2254487

Amit, G., Ben-Ari, R., Hadad, O., Monovich, E., Granot, N., and Hashoul, S. (2017). "Classification of breast MRI lesions using small-size training sets: comparison of deep learning approaches," in *Medical Imaging 2017: Computer-Aided Diagnosis*, eds S. G. Armato III and N. A. Petrick (SPIE), 374–379. doi: 10.1117/12.2249981

Angulakshmi, M., and Lakshmi Priya, G. (2017). Automated brain tumour segmentation techniques—a review. *Int. J. Imaging Syst. Technol.* 27, 66–77. doi: 10.1002/ima.22211

Asif, U., Bennamoun, M., and Sohel, F. A. (2018). A multi-modal, discriminative and spatially invariant CNN for RGB-D object labeling. *IEEE Trans. Patt. Anal. Mach. Intell.* 40, 2051–2065. doi: 10.1109/TPAMI.2017.2747134

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Arch.* Available online at: https://wiki.cancerimagingarchive.net/display/DOI/Segmentation+Labels+and+Radiomic+Features+for+the+Pre-operative+Scans+of+the+TCGA-GBM+collection

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. *Cancer Imaging Arch.* Available online at: https://wiki.cancerimagingarchive.net/display/DOI/Segmentation+Labels+and+Radiomic+Features+for+the+Pre-operative+Scans+of+the+TCGA-GBM+collection

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma MRI collections

with expert segmentation labels and radiomic features. *Sci. Data* 4, 1–13. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *CoRR* abs/1811.02629.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE TPAMI* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Benson, E., Pound, M. P., French, A. P., Jackson, A. S., and Pridmore, T. P. (2018). "Deep hourglass for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 419–428.

Carver, E., Liu, C., Zong, W., Dai, Z., Snyder, J. M., Lee, J., et al. (2018). "Automatic brain tumor segmentation and overall survival prediction using machine learning algorithms," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 406–418.

Castro, E., Cardoso, J. S., and Pereira, J. C. (2018). "Elastic deformations for data augmentation in breast cancer mass detection," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 230–234. doi: 10.1109/BHI.2018.8333411

Chaitanya, K., Karani, N., Baumgartner, C. F., Becker, A., Donati, O., and Konukoglu, E. (2019). "Semi-supervised and task-driven data augmentation," in *Information Processing in Medical Imaging*, eds A. C. S. Chung, J. C. Gee, P. A. Yushkevich, and S. Bao (Cham: Springer International Publishing), 29–41.

Chandra, S., Vakalopoulou, M., Fidon, L., Battistella, E., Estienne, T., Sun, R., et al. (2018). "Context aware 3D CNNs for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 299–310.

Crimi, A., Bakas, S., Kuijf, H. J., Keyvan, F., Reyes, M., and van Walsum, T. (eds). (2019). *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, volume 11384 of Lecture Notes in Computer Science* (Cham).

Dai, L., Li, T., Shu, H., Zhong, L., Shen, H., and Zhu, H. (2018). "Automatic brain tumor segmentation with domain adaptation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 380–392.

Dvornik, N., Mairal, J., and Schmid, C. (2018). On the importance of visual context for data augmentation in scene understanding. *CoRR* abs/1809.02492.

Eaton-Rosen, Z., Bragman, F., Ourselin, S., and Cardoso, M. J. (2019). Improving data augmentation for medical image segmentation. *OpenReview*.

Feng, X., Tustison, N. J., and Meyer, C. H. (2018). "Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 279–288.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing* 321, 321–331. doi: 10.1016/j.neucom.2018.09.013

Fyllingen, E. H., Stensjøen, A. L., Berntsen, E. M., Solheim, O., and Reinertsen, I. (2016). Glioblastoma segmentation: comparison of three different software packages. *PLoS ONE* 11:e0164891. doi: 10.1371/journal.pone.0164891

Galdran, A., Alvarez-Gila, A., Meyer, M. I., Saratxaga, C. L., Araujo, T., Garrote, E., et al. (2017). Data-driven color augmentation techniques for deep skin image analysis. *CoRR* abs/1703.03702.

Gholami, A., Subramanian, S., Shenoy, V., Himthani, N., Yue, X., Zhao, S., et al. (2018). "A novel domain adaptation framework for medical image segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 289–298.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., et al. (2018). NiftyNet: a deep-learning platform for medical imaging. *Comput. Methods Prog. Biomed.* 158, 113–122. doi: 10.1016/j.cmpb.2018.01.025

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (Curran Associates, Inc.), 2672–2680. Available online at: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Gu, S., Meng, X., Sciurba, F. C., Ma, H., Leader, J., Kaminski, N., et al. (2014). Bidirectional elastic image registration using b-spline affine transformation. *Comput. Med. Imaging Graph.* 38, 306–314. doi: 10.1016/j.compmedimag.2014.01.002

Han, C., Murao, K., Satoh, S., and Nakayama, H. (2019). Learning more with less: gan-based medical image augmentation. *CoRR* abs/1904.00838. doi: 10.1145/3357384.3357890

Hollingworth, W., Medina, L. S., Lenkinski, R. E., Shibata, D. K., Bernal, B., Zurakowski, D., et al. (2006). Interrater reliability in assessing quality of diagnostic accuracy studies using the quadas tool: a preliminary assessment. *Acad. Radiol.* 13, 803–810. doi: 10.1016/j.acra.2006.03.008

Huang, Z., and Cohen, F. S. (1996). Affine-invariant b-spline moments for curve matching. *IEEE Trans. Image Process.* 5, 1473–1480. doi: 10.1109/83.536895

Hussain, Z., Gimenez, F., Yi, D., and Rubin, D. (2017). "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA 2017, American Medical Informatics Association Annual Symposium* (Washington, DC).

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 234–244.

Isin, A., Direkoglu, C., and Sah, M. (2016). Review of mri-based brain tumor image segmentation using deep learning methods. *Proc. Comput. Sci.* 102, 317–324. doi: 10.1016/j.procs.2016.09.407

Kao, P., Ngo, T., Zhang, A., Chen, J. W., and Manjunath, B. S. (2018). "Brain tumor segmentation and tractographic feature extraction from structural MR images for overall survival prediction," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 128–141.

Kermi, A., Mahmoudi, I., and Khadir, M. T. (2018). "Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal MRI volumes," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 37–48.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Lachinov, D., Vasiliev, E., and Turlapov, V. (2018). "Glioma segmentation with cascaded unet," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 189–198.

LeCun, Y., Bengio, Y., and Hinton, G. (2016). Deep learning. *Nature* 521, 436–555. doi: 10.1038/nature14539

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, Y., Stojadinovic, S., Hrycushko, B., Wardak, Z., Lau, S., Lu, W., et al. (2017). A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. *PLoS ONE* 12:e0185844. doi: 10.1371/journal.pone.0185844

Ma, J., and Yang, X. (2018). "Automatic brain tumor segmentation by exploring the multi-modality complementary information and cascaded 3d lightweight cNNs," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 25–36.

Marcinkiewicz, M., Nalepa, J., Lorenzo, P. R., Dudzik, W., and Mrukwa, G. (2018). "Segmenting brain tumors from MRI using cascaded multi-modal U-Nets," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 13–24.

McKinley, R., Meier, R., and Wiest, R. (2018). "Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 456–465.

Mehta, R., and Arbel, T. (2018). "3D U-Net for brain tumour segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 254–266.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BraTS). *IEEE TMI* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Mok, T. C. W., and Chung, A. C. S. (2018). Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. *CoRR* abs/1805.11291:1–10.

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 311–320.

Nalepa, J., Lorenzo, P. R., Marcinkiewicz, M., Bobek-Billewicz, B., Wawrzyniak, P., Walczak, M., et al. (2019). Fully-automated deep learning-powered system for DCE-MRI analysis of brain tumors. *CoRR* abs/1907.08303. doi: 10.1016/j.artmed.2019.101769

Nalepa, J., Mrukwa, G., Piechaczek, S., Lorenzo, P. R., Marcinkiewicz, M., Bobek-Billewicz, B., et al. (2019a). "Data augmentation via image registration," in *2019 IEEE International Conference on Image Processing (ICIP)*, 4250–4254.

Nalepa, J., Myller, M., and Kawulok, M. (2019b). Training- and test-time data augmentation for hyperspectral image segmentation. *IEEE Geosci. Remote Sens. Lett.* 1–5. doi: 10.1109/LGRS.2019.2921011

Nguyen, K. P., Fatt, C. C., Treacher, A., Mellema, C., Trivedi, M. H., and Montillo, A. (2019). Anatomically-informed data augmentation for functional MRI with applications to deep learning. *CoRR* abs/1910.08112.

Nuechterlein, N., and Mehta, S. (2018). "3D-ESPNet with pyramidal refinement for volumetric brain tumor image segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 245–253.

Oksuz, I., Ruijsink, B., Puyol-Antón, E., Clough, J. R., Cruz, G., Bustin, A., et al. (2019). Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning. *Med. Image Anal.* 55, 136–147. doi: 10.1016/j.media.2019.04.009

Park, S.-C., Cha, J. H., Lee, S., Jang, W., Lee, C. S., and Lee, J. K. (2019). Deep learning-based deep brain stimulation targeting and clinical applications. *Front. Neurosci.* 13:1128. doi: 10.3389/fnins.2019.01128

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural nets in MRI images. *IEEE TMI* 35, 1240–1251. doi: 10.1109/TMI.2016.2538465

Puybareau, É., Tochon, G., Chazalon, J., and Fabrizio, J. (2018). "Segmentation of gliomas and prediction of patient overall survival: a simple and fast procedure," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 199–209.

Rezaei, M., Harmuth, K., Gierke, W., Kellermeier, T., Fischer, M., Yang, H., et al. (2017). Conditional adversarial network for semantic segmentation of brain tumor. *CoRR* abs/1708.05227:1–10.

Rezaei, M., Yang, H., and Meinel, C. (2018). "voxel-gan: adversarial framework for learning imbalanced brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 321–333.

Ribalta Lorenzo, P., Nalepa, J., Bobek-Billewicz, B., Wawrzyniak, P., Mrukwa, G., Kawulok, M., et al. (2019). Segmenting brain tumors from flair MRI using fully convolutional neural networks. *Comput. Methods Prog. Biomed.* 176, 135–148. doi: 10.1016/j.cmpb.2019.05.006

Rozsa, A., Günther, M., and Boult, T. E. (2016). Towards robust deep neural networks with BANG. *CoRR* abs/1612.00138.

Sahnoun, M., Kallel, F., Dammak, M., Mhiri, C., Ben Mahfoudh, K., and Ben Hamida, A. (2018). "A comparative study of MRI contrast enhancement techniques based on Traditional Gamma Correction and Adaptive Gamma Correction: Case of multiple sclerosis pathology," in *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1–7.

Sauwen, N., Acou, M., Sima, D. M., Veraart, J., Maes, F., Himmelreich, U., et al. (2017). Semi-automated brain tumor segmentation on multi-parametric MRI using regularized non-negative matrix factorization. *BMC Med. Imaging* 17:29. doi: 10.1186/s12880-017-0198-4

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., et al. (2018). "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging*, eds A. Gooya, O. Goksel, I. Oguz, and N. Burgos (Cham: Springer), 1–11.

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6:60. doi: 10.1186/s40537-019-0197-0

Sun, L., Zhang, S., and Luo, L. (2018). "Tumor segmentation and survival prediction in glioma with deep learning," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 83–93.

Tustison, N. J., and Avants, B. B. (2013). Explicit B-spline regularization in diffeomorphic image registration. *Front. Neuroinformatics* 7:39. doi: 10.3389/fninf.2013.00039

Tustison, N. J., Avants, B. B., and Gee, J. C. (2009). Directly manipulated free-form deformation image registration. *IEEE TIP* 18, 624–635. doi: 10.1109/TIP.2008.2010072

Tward, D., and Miller, M. (2017). "Unbiased diffeomorphic mapping of longitudinal data with simultaneous subject specific template estimation," in *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, eds M. J.Cardoso, T.Arbel, E. Ferrante, X. Pennec, A. V. Dalca, S. Parisot, S. Joshi, N. K. Batmanghelich, A. Sotiras, M. Nielsen, M. R. Sabuncu, T. Fletcher, L. Shen, S. Durrleman, and S. Sommer (Cham: Springer International Publishing), 125–136.

Visser, M., Müller, D. M. J., van Duijn, R. J. M., Smits, M., Verburg, N., Hendriks, E. J., et al. (2019). Inter-rater agreement in glioma segmentations on longitudinal MRI. *NeuroImage* 22:101727. doi: 10.1016/j.nicl.2019.101727

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338, 34–45. doi: 10.1016/j.neucom.2019.01.103

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2018). "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, vol. 11384 of Lecture Notes in Computer Science* (Cham), 61–72.

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., and Wang, F. (2017). Generative adversarial networks: introduction and outlook. *IEEE/CAA J. Automat. Sin.* 4, 588–598. doi: 10.1109/JAS.2017.7510583

Wei, W., Liu, L., Truex, S., Yu, L., and Gursoy, M. E. (2018). Adversarial examples in deep learning: characterization and divergence. *CoRR* abs/1807.00051.

Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. (2016). "Understanding data augmentation for classification: when to warp?" in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–6. doi: 10.1109/DICTA.2016.7797091

Yu, B., Zhou, L., Wang, L., Fripp, J., and Bourgeat, P. (2018). "3D cGAN based cross-modality MR image synthesis for brain tumor segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 626–630. doi: 10.1109/ISBI.2018.8363653

Zhang, H., Cissé, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: beyond empirical risk minimization. *CoRR* abs/1710.09412.

Zhao, J., Meng, Z., Wei, L., Sun, C., Zou, Q., and Su, R. (2019). Supervised brain tumor segmentation based on gradient and context-sensitive features. *Front. Neurosci.* 13:144. doi: 10.3389/fnins.2019.00144

Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR* abs/1703.10593. doi: 10.1109/ICCV.2017.244

Conflict of Interest: JN was employed by Future Processing, and MM was employed by Netguru.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 12 December 2019
doi: 10.3389/fncom.2019.00081

Multivariate Analysis of Preoperative Magnetic Resonance Imaging Reveals Transcriptomic Classification of *de novo* Glioblastoma Patients

*Saima Rathore[1,2], Hamed Akbari[1,2], Spyridon Bakas[1,2,3], Jared M. Pisapia[1,4], Gaurav Shukla[1,5], Jeffrey D. Rudie[2], Xiao Da[6], Ramana V. Davuluri[7], Nadia Dahmane[8], Donald M. O'Rourke[9] and Christos Davatzikos[1,2]**

[1] Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [2] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [3] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [4] Division of Neurosurgery, Children Hospital of Philadelphia, Philadelphia, PA, United States, [5] Christiana Care Health System, Philadelphia, PA, United States, [6] Brigham and Women's Hospital, Boston, MA, United States, [7] Department of Biomedical Informatics, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, [8] Department of Neurological Surgery, Weill Cornell Medicine, New York, NY, United States, [9] Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

OPEN ACCESS

Edited by:
Abdelmalik Moujahid, University of the Basque Country, Spain

Reviewed by:
Muhammad Tahir, Saudi Electronic University, Saudi Arabia
Madhura Ingalhalikar, Symbiosis International University, India

***Correspondence:**
Christos Davatzikos
christos.davatzikos@pennmedicine.upenn.edu

Received: *06 July 2019*
Accepted: *12 November 2019*
Published: *12 December 2019*

Citation:
Rathore S, Akbari H, Bakas S, Pisapia JM, Shukla G, Rudie JD, Da X, Davuluri RV, Dahmane N, O'Rourke DM and Davatzikos C (2019) Multivariate Analysis of Preoperative Magnetic Resonance Imaging Reveals Transcriptomic Classification of de novo Glioblastoma Patients. Front. Comput. Neurosci. 13:81. doi: 10.3389/fncom.2019.00081

Glioblastoma, the most frequent primary malignant brain neoplasm, is genetically diverse and classified into four transcriptomic subtypes, i. e., classical, mesenchymal, proneural, and neural. Currently, detection of transcriptomic subtype is based on *ex vivo* analysis of tissue that does not capture the spatial tumor heterogeneity. In view of accumulative evidence of *in vivo* imaging signatures summarizing molecular features of cancer, this study seeks robust non-invasive radiographic markers of transcriptomic classification of glioblastoma, based solely on routine clinically-acquired imaging sequences. A pre-operative retrospective cohort of 112 pathology-proven *de novo* glioblastoma patients, having multi-parametric MRI (T1, T1-Gd, T2, T2-FLAIR), collected from the Hospital of the University of Pennsylvania were included. Following tumor segmentation into distinct radiographic sub-regions, diverse imaging features were extracted and support vector machines were employed to multivariately integrate these features and derive an imaging signature of transcriptomic subtype. Extracted features included intensity distributions, volume, morphology, statistics, tumors' anatomical location, and texture descriptors for each tumor sub-region. The derived signature was evaluated against the transcriptomic subtype of surgically-resected tissue specimens, using a 5-fold cross-validation method and a receiver-operating-characteristics analysis. The proposed model was 71% accurate in distinguishing among the four transcriptomic subtypes. The accuracy (sensitivity/specificity) for distinguishing each subtype (classical, mesenchymal, proneural, neural) from the rest was equal to 88.4% (71.4/92.3), 75.9% (83.9/72.8), 82.1% (73.1/84.9), and 75.9% (79.4/74.4), respectively. The findings were also replicated in The Cancer Genomic Atlas glioblastoma dataset. The obtained imaging signature for the classical subtype was dominated by associations with features related to edge

sharpness, whereas for the mesenchymal subtype had more pronounced presence of higher T2 and T2-FLAIR signal in edema, and higher volume of enhancing tumor and edema. The proneural and neural subtypes were characterized by the lower T1-Gd signal in enhancing tumor and higher T2-FLAIR signal in edema, respectively. Our results indicate that quantitative multivariate analysis of features extracted from clinically-acquired MRI may provide a radiographic biomarker of the transcriptomic profile of glioblastoma. Importantly our findings can be influential in surgical decision-making, treatment planning, and assessment of inoperable tumors.

Keywords: transcriptomic classification, glioblastoma, multivariate analysis, brain tumors classification, biomarkers

INTRODUCTION

Glioblastoma is the most frequent primary malignant brain tumor with grim prognosis, despite aggressive combination of therapies (Stupp et al., 2017), and is characterized by inter- and intra-patient heterogeneity at radiographic, histologic, and molecular fronts, thereby providing opportunities for sub-classification, prognostication, and adoption of targeted therapeutic approaches (Aum et al., 2014; Lemée et al., 2015).

There is mounting evidence that different glioblastoma patients show different levels of sensitivity to therapeutic approaches depending on their distinct genetic characterization. It has been suggested earlier that glioblastoma should not be considered a single disease, but rather should be categorized into four transcriptomic subtypes, i.e., classical, mesenchymal, proneural, and neural (Verhaak et al., 2010). These subtypes present very distinct molecular biomarkers such as collective loss in chromosome 10 and amplification of chromosome 7 in classical subtype, largest occurrence of focal hemizygous deletions of a region at 17q11.2, encompassing *NF1 gene,* in mesenchymal subtype, aberrations in PDGFRA and mutations in *IDH1* in proneural subtype, and presence of *GABRA1, SYT1, NEFL,* and *SLC12A5* in neural subtype (Verhaak et al., 2010). In a recent study by Park et al. it has been shown that subtype-specific genetic aberrations have potential to serve as predictive markers and therapeutic targets (Park et al., 2019).

The determination of the molecular profile of the tumors leads to personalized diagnosis and treatment, as different treatment options may be considered depending on the characteristics of each subtype (Phillips et al., 2006; Verhaak et al., 2010; Bhat et al., 2011). Up until now, the assessment of transcriptomic subtypes was done via molecular profiling of surgical or biopsy tissue. However, such assessment has inherent limitations of: (i) tissue sampling error that sometimes leads to missing the tumor mutation, and (ii) inability to acquire multiple specimens over the course of the disease due to invasiveness of the tissue collection procedure, thereby leading to the failure in determining molecular subtype of the tumor over the course of the treatment.

Analysis of multi-parametric magnetic resonance imaging (mpMRI) data via advanced pattern analytics methods has been progressively shown to provide rich classifications of glioblastoma and its surrounding brain tissue, and has helped identifying relationships between MRI biomarkers and transcriptomic subtypes in gliomas (Gutman et al., 2013; Naeini et al., 2013; Gevaert et al., 2014; Pisapia et al., 2015; Macyszyn et al., 2016; Khened et al., 2019). For instance, the proneural subtype has shown lower levels of contrast enhancement; the mesenchymal subtype has presented lower levels of non-enhanced tumors and intensity in peritumoral edema region (Gutman et al., 2013); the classical subtype has associated necrosis and sharped edges of the edema region (Gevaert et al., 2014). A model to predict the mesenchymal subtype was also proposed (Naeini et al., 2013).

However, this classification scheme has been difficult to translate into clinical practice due to several complicating factors. First, existing literature has found associations between imaging features and individual subtypes (Naeini et al., 2013). Second, most studies to date have used basic imaging sequences only or have used very few hand-crafted imaging features, failing to leverage the power of computationally extracting and selecting imaging features (recently called radiomics), and analyzing them through advanced pattern analysis methods to build a more powerful predictive model (Gutman et al., 2013; Gevaert et al., 2014; Macyszyn et al., 2016). As literature increasingly acknowledges the tumor spatial and temporal heterogeneity, there is a parallel focus on extracting extensive features of the tumor and its surrounding peritumoral region toward providing a better characterization of patients. Furthermore, analysis of advanced mpMRI data can provide more details, which might not be available in conventional imaging.

This study aims to determine the transcriptomic subtypes of *de novo* glioblastoma patients by multivariately assessing imaging features from routine clinically-acquired scans, reflecting tumor biological properties such as angiogenesis, proliferation, cellularity, and peritumoral infiltration. Identifying these transcriptomic subtypes may allow enrollment of patients into targeted clinical trials, longitudinal profiling of the tumor, and assessment of treatment response.

MATERIALS AND METHODS

Study Setting and Data Source

This study evaluates a group of 112 primary glioblastoma patients, diagnosed between 2006 and 2013 at the Hospital of the University of Pennsylvania (HUP), having pre-operative mpMRI

(T1, T2, T1-Gd, T2-FLAIR). A subset of these patients (n = 89) had additional diffusion tensor imaging (DTI) and dynamic susceptibility contrast-enhanced (DSC) MRI imaging available. The proposed classification models were developed on n =112 patients using conventional imaging only (T1, T2, T1-Gd, T2-FLAIR), whereas the subset of the patients (n = 89) was used to further analyze the imaging properties of different subtypes. The overall analysis was carried out on HUP dataset, and the findings were then replicated independently in The Cancer Genomic Atlas glioblastoma (TCGA-GBM) dataset (Clark et al., 2013; Scarpace et al., 2013) (n = 60), part of the International Brain Tumor Segmentation (BraTS) challenge dataset (Menze et al., 2015; Bakas et al., 2018), having the same set of pre-operative mpMRI. Expert manual segmentations for this dataset were downloaded from The Cancer Imaging Archive (TCIA) website (Bakas et al., 2017a,b). The study population was uniformly distributed and did not have any statistically significant difference based on clinical and demographical factors. All experiments were approved by the Institutional Review Board (IRB) of the University of Pennsylvania (approval no: 706564) and written informed consent was obtained from all patients. All experiments were carried out in accordance with the guidelines and regulations of the approved IRB.

Transcriptomic Subtyping

After pathologic confirmation of glioblastoma diagnosis, all tumors underwent subtyping into one of the four transcriptomic subtypes (classical = 21, mesenchymal = 31, proneural = 26, neural = 34). For this subtyping, we used an isoform-level assay classifier initially constructed using exon array data from glioblastoma samples in TCIA (Verhaak et al., 2010). It was then translated into a clinically applicable platform, where expression of desired transcripts was measured using reverse transcriptase–quantitative polymerase chain reaction (RT-qPCR) (Pal et al., 2014). RNA was isolated from the tissue samples using Tri Reagent (Sigma). A high-capacity complementary DNA reverse transcriptase kit (Applied Biosystems) was used to reverse-transcribe the RNA, and qPCR was then performed to designate the subtype. The assay was based on the expression of 121 transcripts with four housekeeping genes as controls.

Pre-processing Applied on the Dataset

All MRI of each patient were pre-processed using a series of image processing steps, including: (i) smoothing (i.e., reducing high frequency noise variations while preserving underlying anatomical structures) using Smallest Univalue Segment Assimilating Nucleus (SUSAN) denoising (Smith et al., 1997); (ii) correction for magnetic field inhomogeneity using N3 bias correction (Tustison et al., 2010); (iii) co-registration of all MRIs of each patient at 12-degrees of freedom for examining anatomically aligned signals at the voxel level using affine registration through the Linear Image Registration Tool (Jenkinson and Smith, 2001); (iv) skull stripping using the Brain Extraction Tool (Smith, 2002); and (v) matching of intensity profiles (histogram matching) of all MRIs of all patients to the corresponding MRIs of a reference patient.

Following the pre-processing, all tumors were segmented in distinct radiographic sub-regions of peritumoral edema region (ED), enhancing tumor (ET), and non-enhancing tumor (NET) (**Figure 1**) using a computational algorithm [namely GLISTRboost (Gooya et al., 2011; Bakas et al., 2016)]. The segmentations were assessed by two expert readers (H.A., G.S.) and revised before image analysis, when necessary. The segmentations were transformed into a standard atlas space to produce a standardized statistical distribution atlas for quantifying the tumor spatial location.

Radiophenotypic Tumor Characterization

The radiophenotypic characteristics of each tumor were quantified using a comprehensive and diverse set of imaging features, extracted from all tumor sub-regions (i.e., ED, ET, NET) and all MRI sequences using the Cancer Imaging Phenomics Toolkit (CaPTk) (Davatzikos et al., 2018). The feature set extracted to build the predictive model for this study comprised of (i) volumetric measurements, (ii) morphology parameters, (iii) location information, and (iv) statistical moments of the intensity distributions. The volumetric, location, and intensity statistics were calculated in 3D. The volumetric measurements include volume and surface area measurements of ED, NET, ET, tumor core (TC), which is the union of NET and ET, and whole tumor (WT), which is the combination of TC with ED. In addition, ratios of the volumes of the various tumor sub-regions and their union over the brain volume, were also calculated.

To capture the spatial distribution of each tumor, eight spatial distribution atlases were constructed as introduced in Akbari et al. (2018), two for each molecular subtype, i.e., $P^{(+)}$ and $P^{(-)}$ for proneural and non-proneural tumors, respectively. These distribution atlases were generated by superimposing the TC (ET+NET) segmentation labels of all patients according to their transcriptomic subtype status, i.e., superimposing the TC labels of proneural and non-proneural tumors. The similarity of the distribution pattern of an unseen tumor is then calculated by considering the intersection area between the tumor and the spatial map (**Figure S1**). Maximum and average frequency for each spatial distribution atlas in the intersected area are estimated, and four discrete relative values (L1, L2, L3, and L4) are used to evaluate any new unseen patient, for each subtype, thereby leading to a total of 16 location features.

$$
\begin{aligned}
L_1 &= mean\left[P^+\right] - mean[\,P^-], \\
L_2 &= \max\left[P^+\right] - max\left[P^-\,\right], \\
L_3 &= \frac{mean\left[P^+\right]}{mean\left[P^-\,\right]}, \\
L_4 &= \frac{\max\left[P^+\right]}{\max\left[P^-\right]}
\end{aligned}
$$

Moreover, the distance of various tumor sub-regions, e.g., ED, TC, from the ventricles, and the proportions of TC in each lobe of the brain have also been utilized as additional location features. The proportion of TC in various brain regions,

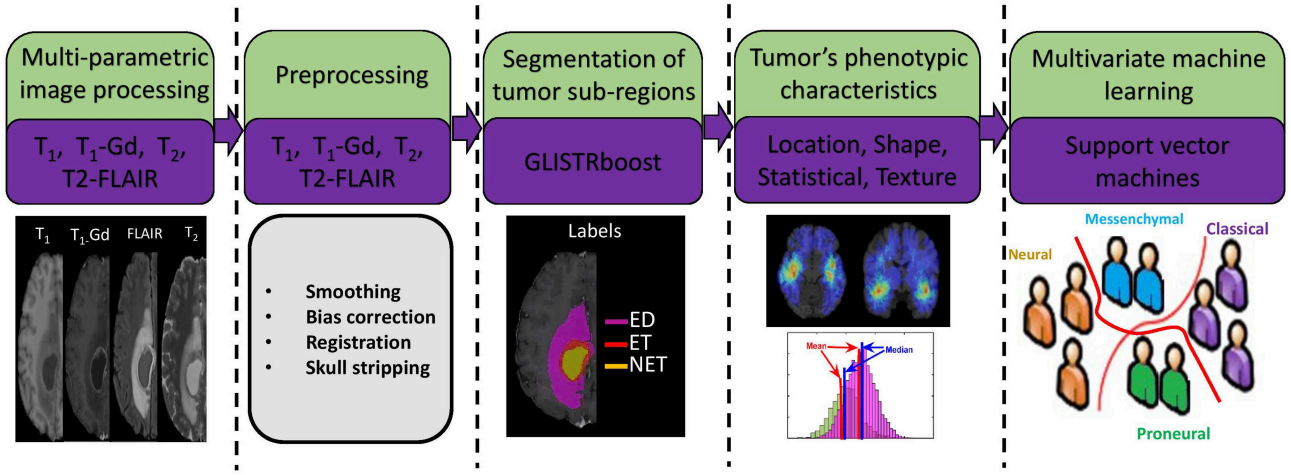


FIGURE 1 | Image post-processing workflow. I. Preoperative imaging sequences (examples: T1-Gd, T2-FLAIR, T1, T2). II. Preprocessing. III. Segmentation of various subregions of tumor such as peritumoral edema region (ED), non-enhancing tumor core (NET), and enhancing tumor (ET). IV. Extraction of radiomic features in the segmented regions (ED, ET, NET) using all the imaging sequences. V. Multivariate machine learning model of support vector machines with 5-fold cross validation and sequential feature selection.

including temporal, frontal, parietal, occipital, basal ganglia, cc fornix, insula, cerebellum, and brain stem, was calculated by mapping each image to an atlas template via a deformable registration method (Gooya et al., 2011) that not only accounts for mass effect but also takes care of inter-individual anatomical variations (Kwon et al., 2014).

Furthermore, we used first-order statistical moments of intensity distributions to quantify the phenotypic characteristics of each tumor sub-region, along with second-order statistics that describe textural properties in tumor sub-regions. A gray-level co-occurrence matrix was calculated by considering the voxels within a radius of 1 and in the 13 main directions, and texture features of contrast, correlation, energy, and homogeneity were extracted. The intensity profiles of various sub-regions of tumor were also quantified using histograms. These histograms are reflective of the changes caused by the tumor both at functional and anatomical levels, which in turn change the corresponding imaging signals, and have shown strong association with various outcome of interest (Macyszyn et al., 2016; Rathore et al., 2018). Here, each intensity distribution is divided in to 5 bins and percentage of voxels in each bin are calculated.

Morphology parameters, comprising area, perimeter, extent, solidity, and length of major- and minor-axis, were extracted from one 2D slice per tumor. In order to pick the 2D slice for extraction of morphological features, we traversed the image in the axial direction and found the slice that had largest area of tumor core.

Feature Selection and Predictive Model Development

Support Vector Machines (SVM) (Chang and Lin, 2011), that has been extensively used in the past in medical image classification/segmentation (Lao et al., 2008; Haller et al., 2013), was used for predictive modeling in this study. We dealt the problem of classification as 4 one-vs. -rest classification problems. We trained a separate SVM to discriminate between one transriptomic subtype and the rest of the subtypes, such as classical vs. others, mesenchymal vs. others, neural vs. others, and proneural vs. others. To confirm the robustness of the method and to ensure that estimates of accuracy would be likely to generalize to new patients, we evaluated all classifiers through 5-fold cross-validation. In each iteration of the cross-validation, feature selection and classifier's parameters optimization was performed on the training folds and the resulting classification model, developed solely on the training folds, was applied on new/unseen test fold. Sequential forward feature selection was employed at each iteration until convergence, i.e., there was no improvement over a specific threshold. The final classification performance was obtained by combining the predictions of individual classifiers. For each classifier, the particular subtype was considered positive class and the rest of the subtypes were considered negative class. The distance of the sample from the hyperplane was noted for each classifier and highest distance was chosen as the final label of the sample. For example, if proneural, neural, mesenchymal, and classical have 0.45, 3.54, −2.43, and 5.32, then the classical label was assigned to the sample.

The classification performance of the proposed models was evaluated in terms of accuracy, balanced accuracy, sensitivity, and specificity. Sensitivity and specificity refer to the percentage of correctly classified samples of positive and negative classes, respectively. Balanced accuracy is the average of the proportion corrects of each class individually, whereas accuracy is the total proportion corrects of the population.

Statistical Analysis

The statistical analysis was performed with R (version 3.3.2, http://www.R-project.org), SPSS (version 25.0.0.0, IBM), and MatLab (version R2014b, Mathworks). For evaluation of statistically significant imaging features associated with each subtype, we used Kruskal-Wallis test (Chan et al., 1997).

TABLE 1 | Performance of the proposed transcriptomic subtype prediction model in terms of various performance measures.

Classification performance	Imaging subtypes *(n)*				
	Proneural (*n* = 26)	**Neural (*n* = 34)**	**Mesenchymal (*n* = 31)**	**Classical (*n* = 21)**	**Overall (*n* = 112)**
Accuracy	82.14	75.89	75.89	88.39	71.00
Balanced accuracy	78.98	76.89	78.36	81.87	79.01
Sensitivity	73.08	79.41	83.87	71.43	77.68
Specificity	84.88	74.36	72.84	92.31	79.75
AUC	0.82	0.78	0.81	0.84	---

First four rows show the result for binary classification wherein each subtype is classified against the rest of subtypes. The last row shows the final 4-way classification accuracy obtained by combining the predictions of individual classifiers.

RESULTS

Performance of the Transcriptomic Subtype Prediction Model

The cross-validated accuracy and balanced accuracy [BA] of the obtained classifiers for classical, mesenchymal, proneural, and neural subtypes was 88.4% [BA: 81.9%], 75.9% [BA: 78.4%], 82.1% [BA: 78.9%], and 75.9% [BA: 76.9%], respectively. The overall 4-way classification among the four transcriptomic subtypes was 71% under 5-fold cross-validation experiment. Performance of the proposed prediction model is given in **Table 1** where the first four columns show the result for binary classification, wherein each transcriptomic subtype is classified against the rest of the subtypes, and the last column shows the final 4-way classification accuracy obtained by combining the predictions of individual classifiers.

Receiver-operating-characteristic (ROC) analysis on the given dataset yielded an area-under the-curve (AUC) of 0.82, 0.78, 0.81, and 0.84, for proneural, neural, mesenchymal, and classical subtypes, respectively (**Figure 2**).

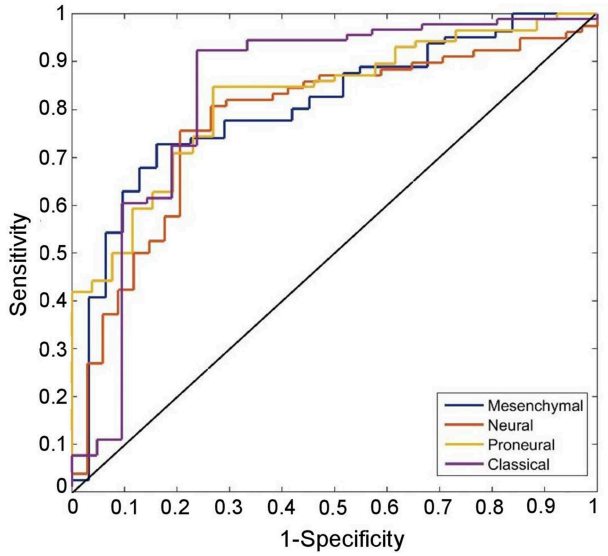


FIGURE 2 | ROC curves of the predicted transriptomic subtypes are compared with chance (the diagonal line). ROC curves correctly classify proneural, neural, mesenchymal, and classical subtypes with 82.1% (sensitivity: 73.1, specificity = 84.9), 75.9% (sensitivity: 79.4, specificity = 74.4), 75.9% (sensitivity: 83.9, specificity = 72.8), and 88.4% (sensitivity: 71.4, specificity = 92.3) classification success rate, respectively.

Important Phenotypic Characteristics of Different Transcriptomic Subtypes

Along with evaluating the predictive performance of the model, we assessed individual features with the most predictive value. Our results have shown that specific subtypes have quite distinct quantitative imaging features, which can be utilized (**Table 2**, **Figure 3**). The main characteristics of the obtained imaging signature show that the mesenchymal subtype (in comparison with other subtypes) have lower T2 and T2-FLAIR signal in peritumoral edematous/invaded region, ET of lower eccentricity, NET of higher eccentricity, and higher volumes of ET, ED and WT. The proneural subtype, compared with the other subtypes, included signals of lower and uniform T1-Gd in ET. The neural subtype showed signals of higher T2-FLAIR in ED and lower eccentricity of NET, and the classical subtype showed smaller surface area of ED and WT.

TABLE 2 | Important imaging characteristics that distinguish each subtype from the rest of the subtypes.

Imaging subtypes *(n)*			
Proneural (*n* = 26)	**Neural (*n* = 34)**	**Mesenchymal (*n* = 31)**	**Classical (*n* = 21)**
Lower Signal in ET (T1-Gd)	Higher signal in ED (T2-FLAIR)	Lower signal in ED (T2-FLAIR)	Surface area (ED, WT)
Higher uniformity in ET (T1-Gd)	Lower eccentricity (NET)	Lower signal in ED (T2)	
		Lower eccentricity (ET)	
		Higher eccentricity (NET)	
		Bigger volume (ED, ET, WT)	

Replication of the Proposed Model in TCIA Dataset

The predictive performance of the proposed model was also evaluated in an independent replication dataset of pre-operative glioblastoma patients, downloaded from TCIA (Bakas et al., 2017a), by applying the model trained on the discovery (i.e., HUP) dataset. The information about the molecular subtypes of TCIA patients was acquired from existing studies (Verhaak et al., 2010; Park et al., 2019). The four models, pertaining to four different molecular subtypes, trained on HUP dataset were applied to the patients in the replication (i.e., TCIA) cohort. The final molecular status of each patient in the replication dataset was obtained by combining the predictions of individual

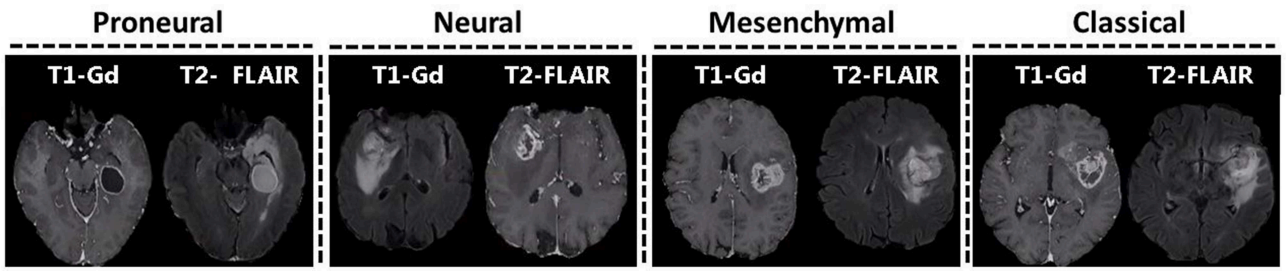


FIGURE 3 | Representative images of different subtypes.

classifiers, as done in the discovery dataset, leading to 69% classification success rate compared to 25% chance in 4-way classification accuracy.

DISCUSSION

We identified an *in vivo* radiographic signature of transcriptomic subtypes in glioblastoma by using quantitative multivariate analysis of mpMRI in a non-invasive manner, and further attempted to provide patho-physiological associations of the most distinctive imaging features. An important existing study has demonstrated the potential that deep learning techniques can be used for identifying associations between brain imaging phenotypes and genomic characteristics (Khened et al., 2019). The hereby proposed method is different from existing literature (Macyszyn et al., 2016; Khened et al., 2019) on the breadth of extracted mpMRI-based features, leading to an extensive radiographic signature. The proposed signature sheds light into the anatomical and pathological characteristics of the tumor, via macroscopic imaging features summarizing tumor characteristics related to water concentration, blood-brain barrier breakage, cell density, uniformity/heterogeneity, and geometric variations. We have achieved these findings utilizing routine mpMRI scans acquired under current clinical practice for glioblastoma, without the need to utilize any molecular imaging methods. We evaluated our model via a cross-validation mechanism in the HUP dataset, and also performed a multi-institutional validation to demonstrate generalizability. Potential applications of this signature include facilitating the assessment of transcriptomic status for patients with inadequate tissue. In a recent study by Park et al., it has been shown that subtype-specific genetic aberrations have potential to serve as predictive markers and therapeutic targets (Park et al., 2019). Therefore, in case of subtype-targeted clinical trials, it becomes very important to distinguish one particular subtype from the rest. The automatic distinction of these subtypes leads to personalized diagnosis and treatment, as different options may be considered depending on the histologic characteristics of different subtypes.

Biological Explanation of Quantitative Features of Different Subtypes

Toward gaining an understanding about the biological developments that induce different mutation status, we analyzed in isolation each individual feature that we used to develop our classification models. The analysis revealed that each subtype had an accompanying distinct and comprehensive set of radiographically relatable features (**Table 2**). The main findings from comparing the features of different transcriptomic subtypes, in ET, NET, and ED, are as follows:

1. Regions of lower and uniform T1-Gd signal in proneural subtype, suggestive of less blood–brain barrier compromise;
2. Areas of lower water content in mesenchymal subtype, reflected by T2-FLAIR and T2-weighted imaging, consistent with the characteristics of dense tissue;
3. Larger surface area of ED and WT in mesenchymal subtype, which points toward deep infiltration and migratory nature of the tumor;
4. Smaller surface area of ED and WT in classical subtype, supporting a radiographic phenotype of compact and less migratory nature of the tumor;
5. Major to minor axes ratios, associated with NET in neural subtype and ET/NET in mesenchymal subtype, were different from other subtypes (**Table 2**). The major axis was characterized by the longest possible 2D distance in a region; minor axis is vertical to the major axis. This eccentricity measure is suggestive of regular/spherical NET in neural subtype and irregular NET in mesenchymal subtype.
6. Regions of relatively lower contrast of T1 imaging sequence in ET in neural subtype, suggestive of more uniform T1 signal (**Table 2**).

It is important to note that despite several discriminative features, neither of these features is sufficient enough to predict transcriptomic subtype on each patient basis. However, synergistic integration of these features via appropriate machine learning yielded reasonable sensitivity and specificity in predicting subtype on an individual patient basis, thereby underscoring the potential of multivariate analysis methods.

Discriminative Power of Advanced MRI (DTI and DSC-MRI) Modalities

Advanced MRI sequences were evaluated to probe their discriminative power, compared to that of structural (conventional) imaging, i.e., T1, T2, T2-Flair, and T1-Gd. It is worth mentioning that these imaging sequences were not

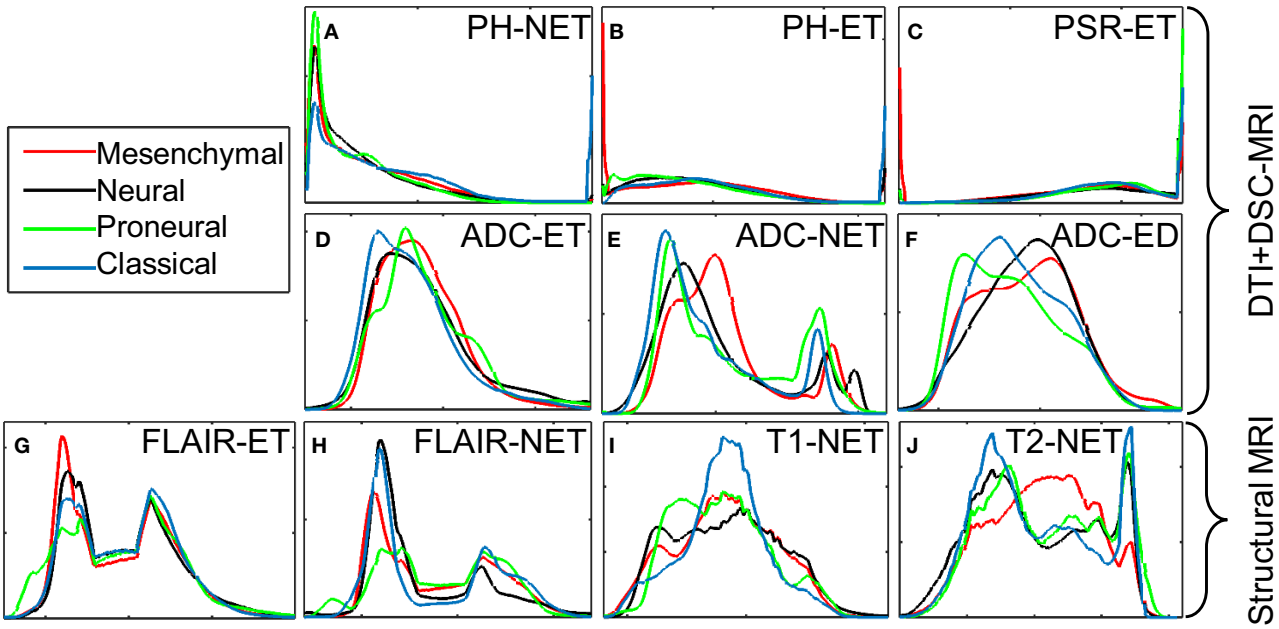


FIGURE 4 | Intensity histograms display the diffusion and perfusion measures of different transcriptomic subtypes as measured by the DTI and DSC-MRI signals. **(A)** PH intensity in NET; **(B)** PH intensity in ET; **(C)** PSR intensity in ET; **(D)** ADC intensity in ET; **(E)** ADC intensity in NET; **(F)** ADC intensity in ED; **(G)** T2-Flair intensity in ET; **(H)** T2-Flair intensity in NET; **(I)** T1 intensity in NET; **(J)** T2 intensity in NET. The measures displayed in the first two rows are for analysis only; these measures have not been used for building the model.

utilized to develop the classification models, rather only to analyze the diffusion and perfusion characteristics of a subset of these patients. These additional sequences comprised derivatives of DTI [i.e., fractional anisotropy (FA), apparent diffusion coefficient (ADC), radial diffusivity (RAD), axial diffusivity (AX)], as well as DSC-MRI derivatives, i.e., percentage signal recovery (PSR), peak height (PH), and relative cerebral blood volume (rCBV).

Imaging derivatives of DTI are reflective of the water diffusion process, which is partially affected by the architecture and density of tumor cells (Lu et al., 2003), in brain. The classical subtype has larger regions of lower ADC determined by the histograms (**Figure 4**) in NET ($p = 2.27 \times 10^{-08}$) and ET ($p = 1.97 \times 10^{-07}$) of the tumor, suggestive of less watery, and denser tumors. Imaging derivatives of DSC-MRI enumerate microvasculature and hemodynamics characteristics of the tumor (Wintermark et al., 2005; Tykocinski et al., 2012). When volume of brain tumors exceeds a certain critical limit, the consequential ischemia activates the discharge of angiogenic factors, which in turn endorses vascular proliferation and eventually leads to the formation of leaky and torturous tumor vessels (Lev and Hochberg, 1998; McDonald and Choyke, 2003; Bullitt et al., 2005; Hicklin et al., 2005; Essock-Burns et al., 2011; Thompson et al., 2011; Swami, 2013; Jensen et al., 2014). These imaging derivatives also steered toward some key findings. The classical subtype showed imaging features in agreement with highly vascular tumor, as shown by the PH in ET ($p = 1.54 \times 10^{-15}$) and NET ($p = 4.00 \times 10^{-06}$), revealing increased and compromised micro-vascularity compared to other subtypes. On the other hand, the proneural subtype had increased PSR in ET, indicative of lower micro-vascularity compared to other subtype.

Clinical Relevance and Impact

The assessment of transcriptomic subtype of glioblastoma via analysis of tissue specimen can be limited due to sampling error, and reluctance for longitudinal assessment of the status due to invasive nature of surgery. Our proposed imaging signature has potential to address both these limitations, since mpMRI facilitates assessment and monitoring of the tumor in its entirety in a repeatable manner. Further, the non-invasive imaging signature captures the heterogeneity of the whole tumor extent, instead of the analysis of one tissue specimen, therefore provides a global perspective of the transcriptomic status of a tumor. Our imaging signature is derivative of mpMRI that is routinely acquired for glioblastoma patients, therefore, is ready for immediate translation to the clinic. While the current method focuses on non-invasive assessment of transcriptomic subtype status, the same approach could also be used for molecular assessment in general. Further, the proposed non-invasive imaging signature can be applied to recurrent glioblastoma, with the goal of determining transcriptomic subtype status before, during, and after the treatment. This would help in non-invasive monitoring of dynamic changes in transcriptomic subtypes as response to targeted therapeutic approaches and consequently would in turn allow for tailoring the adopted therapies.

CONCLUSION

We can quantify important imaging characteristics within various sub-regions of the tumor and detect its transcriptomic subtype only by examining mpMRI data using advanced analytical methods and without the need of advanced genetic testing. The present study extracts an extensive set of quantitative imaging phenomic features from structural MRI sequences, and employs these variables via machine learning techniques to non-invasively distinguish transcriptomic glioblastoma subtypes. This molecular classification, due to its distinct phenotypic pattern derived from routine MRI, renders our imaging signature of increased likelihood for effective and immediate translation into clinical practice. The use of cross-validation within HUP dataset and the replication of our findings on TCIA dataset provide confidence in the generalizability of these subtypes and the proposed method on other datasets.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board (IRB) of University of Pennsylvania (approval no: 706564). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SR, HA, SB, and CD: conceptualization. SR: methodology, software, formal analysis, and writing—original draft. HA, SB, JP, GS, JR, XD, RD, ND, DO'R, and CD: validation. SR and CD: resources. SR, RD, and ND: data curation. SR, HA, SB, JP, GS, JR, XD, RD, ND, DO'R, and CD: writing—review and editing. CD: supervision and funding acquisition.

FUNDING

This research was funded by National Institutes of Health Grants (R01-NS042645, U24-CA189523) and Abramson Cancer Center, Hospital of the University of Pennsylvania Grant (ACC 040-0427-4-572593-xxxx-2433-8348).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00081/full#supplementary-material

REFERENCES

Akbari, H., Bakas, S., Pisapia, J. M., Nasrallah, M. P., Rozycki, M., Martinez-Lage, M., et al. (2018). *In vivo* evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature. *Neuro-Oncology* 20, 1068–1079. doi: 10.1093/neuonc/noy033

Aum, D. J., Kim, D. H., Beaumont, T. L., Leuthardt, E. C., Dunn, G. P., Kim, A., et al. (2014). Molecular and cellular heterogeneity: the hallmark of glioblastoma. *Neurosurg. Focus* 37:E11. doi: 10.3171/2014.9.FOCUS14521

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Arch.* doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scient. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., and Jakab, A. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv:1811.02629. [Preprint].*

Bakas, S., Zeng, K., Sotiras, A., Rathore, S., Akbari, H., Gaonkar, B., et al. (2016). GLISTRboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. *Brainlesion* 9556, 144–155. doi: 10.1007/978-3-319-30858-6_13

Bhat, K. P., Salazar, K. L., Balasubramaniyan, V., Wani, K., Heathcock, L., Hollingsworth, F., et al. (2011). The transcriptional coactivator TAZ regulates mesenchymal differentiation in malignant glioma. *Genes Dev.* 25, 2594–2609. doi: 10.1101/gad.176800.111

Bullitt, E., Zeng, D., Gerig, G., Aylward, S., Joshi, S., Smith, J. K., et al. (2005). Vessel tortuosity and brain tumor malignancy: a blinded study1. *Acad. Radiol.* 12, 1232–1240. doi: 10.1016/j.acra.2005.05.027

Chan, Y., Walmsley, R., and P. (1997). Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Phys. Ther.* 77, 1755–1762. doi: 10.1093/ptj/77.12.1755

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

Davatzikos, C., Rathore, S., Bakas, S., Pati, S., Bergman, M., Kalarot, R., et al. (2018). Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J. Med. Imaging* 5:011018. doi: 10.1117/1.JMI.5.1.011018

Essock-Burns, E., Lupo, J. M., Cha, S., Polley, M.-Y., Butowski, N. A., Chang, S. M., et al. (2011). Assessment of perfusion MRI-derived parameters in evaluating and predicting response to antiangiogenic therapy in patients with newly diagnosed glioblastoma. *Neuro-oncology* 13, 119–131. doi: 10.1093/neuonc/noq143

Gevaert, O., Mitchell, L. A., Achrol, A. S., Xu, J., Echegaray, S., Steinberg, G. K., et al. (2014). Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology* 273, 168–174. doi: 10.1148/radiol.14131731

Gooya, A., Biros, G., and Davatzikos, C. (2011). Deformable registration of glioma images using EM algorithm and diffusion reaction modeling. *IEEE Trans. Med. Imaging* 30, 375–390. doi: 10.1109/TMI.2010.2078833

Gutman, D. A., Cooper, L. A., Hwang, S. N., Holder, C. A., Gao, J., Aurora, T. D., et al. (2013). MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* 267, 560–569. doi: 10.1148/radiol.13120118

Haller, S., Missonnier, P., Herrmann, F., Rodriguez, C., Deiber, M., Nguyen, D., et al. (2013). Individual classification of mild cognitive impairment subtypes by support vector machine analysis of white matter DTI. *Am. J. Neuroradiol.* 34, 283–291. doi: 10.3174/ajnr.A3223

Hicklin, D. J., Ellis, L., and M. (2005). Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. *J. Clin. Oncol.* 23, 1011–1027. doi: 10.1200/JCO.2005.06.081

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156. doi: 10.1016/S1361-8415(01)00036-6

Jensen, R. L., Mumert, M. L., Gillespie, D. L., Kinney, A. Y., Schabel, M. C., Salzman, K., et al. (2014). Preoperative dynamic contrast-enhanced MRI correlates with molecular markers of hypoxia and vascularity in specific areas of intratumoral microenvironment and is predictive of patient outcome. *Neuro-oncology* 16, 280–291. doi: 10.1093/neuonc/not148

Khened, M., Anand, V. K., Acharya, G., Shah, N., and Krishnamurthi, G. (2019). "3D convolution neural networks for molecular subtype prediction in glioblastoma multiforme," in *SPIE Medical Imaging* (San Diego, CA), 10954.

Kwon, D., Shinohara, R. T., Akbari, H., and Davatzikos, C. (2014). "Combining generative models for multifocal glioma segmentation and registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol.* 2014 (Boston, MA), 763–770. doi: 10.1007/978-3-319-10404-1_95

Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., et al. (2008). Computer-assisted segmentation of white matter lesions in 3D MR images, using support vector machine. *Acad. Radiol.* 15, 300–313. doi: 10.1016/j.acra.2007.10.012

Lemée, J.-M., Clavreul, A., and Menei, P. (2015). Intratumoral heterogeneity in glioblastoma: don't forget the peritumoral brain zone. *Neuro Oncol.* 17, 1322–1332. doi: 10.1093/neuonc/nov119

Lev, M. H., and Hochberg, F. (1998). Perfusion magnetic resonance imaging to assess brain tumor responses to new therapies. *Cancer Control* 5, 115–123. doi: 10.1177/107327489800500202

Lu, S., Ahn, D., Johnson, G., and Cha, S. (2003). Peritumoral diffusion tensor imaging of high-grade gliomas and metastatic brain tumors. *Am. J. Neuroradiol.* 24, 937–941.

Macyszyn, L., Akbari, H., Pisapia, J. M., Da, X., Attiah, M., Pigrish, V., et al. (2016). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-oncology* 18, 417–425. doi: 10.1093/neuonc/nov127

McDonald, D. M., Choyke, P. L. (2003). Imaging of angiogenesis: from microscope to clinic. *Nat. Med.* 9, 713–725. doi: 10.1038/nm0603-713

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Naeini, K. M., Pope, W. B., Cloughesy, T. F., Harris, R. J., Lai, A., Eskin, A., et al. (2013). Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images. *Neuro-oncology* 15, 626–634. doi: 10.1093/neuonc/not008

Pal, S., Bi, Y., Macyszyn, L., Showe, L. C., O'Rourke, D. M., Davuluri, R., et al. (2014). Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes. *Nucleic Acids Res.* 42:e64. doi: 10.1093/nar/gku121

Park, A. K., Kim, P., Ballester, L. Y., Esquenazi, Y., and Zhao, Z. (2019). Subtype-specific signaling pathways and genomic aberrations associated with prognosis of glioblastoma. *Neuro-oncology* 21, 59–70. doi: 10.1093/neuonc/noy120

Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., et al. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9, 157–173. doi: 10.1016/j.ccr.2006.02.019

Pisapia, J. M., Macyszyn, L., Akbari, H., Da, X., Pigrish, V., Attiah, M. A., et al. (2015). 135 imaging patterns predict patient survival and molecular subtype in glioblastoma using machine learning techniques. *Neurosurgery* 62(Suppl. 1):209. doi: 10.1227/01.neu.0000467097.06935.d9

Rathore, S., Akbari, H., Rozycki, M., Abdullah, K. G., Nasrallah, M. P., Binder, Z. A., et al. (2018). Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Nat. Scient. Rep.* 8:5087. doi: 10.1038/s41598-018-22739-2

Scarpace, L., Mikkelsen, T., Cha, S., Rao, S., Tekchandani, S., Gutman, D., et al. (2013). Radiology data from the cancer genome atlas glioblastoma multiforme [TCGA-GBM] collection. *Cancer Imaging Arch.* 2016. doi: 10.7937/K9/TCIA.2016.RNYFUYE9

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Smith, S. M., Brady, J., and M. (1997). SUSAN - a new approach to low level image processing. *Int. J. Comput. Vis.* 23, 45–78. doi: 10.1023/A:1007963824710

Stupp, R., Taillibert, S., Kanner, A., Read, W., Steinberg, D., Lhermitte, B., et al. (2017). Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: a randomized clinical trial. *JAMA* 318, 2306–2316. doi: 10.1001/jama.2017.18718

Swami, M. (2013). Cancer: enhancing EGFR targeting. *Nat. Med.* 19, 682–682. doi: 10.1038/nm.3236

Thompson, G., Mills, S., Coope, D., O'connor, J., and Jackson, A. (2011). Imaging biomarkers of angiogenesis and the microvascular environment in cerebral tumours. *Br. J. Radiol.* 84, S127–S144. doi: 10.1259/bjr/66316279

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Tykocinski, E. S., Grant, R. A., Kapoor, G. S., Krejza, J., Bohman, L.-E., Gocke, T. A., et al. (2012). Use of magnetic perfusion-weighted imaging to determine epidermal growth factor receptor variant III expression in glioblastoma. *Neuro-oncology* 14, 613–623. doi: 10.1093/neuonc/nos073

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Wintermark, M., Sesay, M., Barbier, E., Borbély, K., Dillon, W. P., Eastwood, J. D., et al. (2005). Comparative overview of brain perfusion imaging techniques. *J. Neuroradiol.* 32, 294–314. doi: 10.1016/S0150-9861(05)83159-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 20 December 2019
doi: 10.3389/fncom.2019.00084

Multi-Disease Segmentation of Gliomas and White Matter Hyperintensities in the BraTS Data Using a 3D Convolutional Neural Network

Jeffrey D. Rudie[1,2], David A. Weiss[3], Rachit Saluja[4], Andreas M. Rauschecker[2], Jiancong Wang[1], Leo Sugrue[2], Spyridon Bakas[1,5,6†] and John B. Colby[2†]*

[1] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [2] Department of Radiology & Biomedical Imaging, University of California, San Francisco, San Francisco, CA, United States, [3] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, United States, [4] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, United States, [5] Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, United States, [6] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

An important challenge in segmenting real-world biomedical imaging data is the presence of multiple disease processes within individual subjects. Most adults above age 60 exhibit a variable degree of small vessel ischemic disease, as well as chronic infarcts, which will manifest as white matter hyperintensities (WMH) on brain MRIs. Subjects diagnosed with gliomas will also typically exhibit some degree of abnormal T2 signal due to WMH, rather than just due to tumor. We sought to develop a fully automated algorithm to distinguish and quantify these distinct disease processes within individual subjects' brain MRIs. To address this multi-disease problem, we trained a 3D U-Net to distinguish between abnormal signal arising from tumors vs. WMH in the 3D multi-parametric MRI (mpMRI, i.e., native T1-weighted, T1-post-contrast, T2, T2-FLAIR) scans of the International Brain Tumor Segmentation (BraTS) 2018 dataset (n_{training} = 285, $n_{\text{validation}}$ = 66). Our trained neuroradiologist manually annotated WMH on the BraTS training subjects, finding that 69% of subjects had WMH. Our 3D U-Net model had a 4-channel 3D input patch (80 × 80 × 80) from mpMRI, four encoding and decoding layers, and an output of either four [background, active tumor (AT), necrotic core (NCR), peritumoral edematous/infiltrated tissue (ED)] or five classes (adding WMH as the fifth class). For both the four- and five-class output models, the median *Dice* for whole tumor (WT) extent (i.e., union of AT, ED, NCR) was 0.92 in both training and validation sets. Notably, the five-class model achieved significantly (p = 0.002) lower/better Hausdorff distances for WT extent in the training subjects. There was strong positive correlation between manually segmented and predicted volumes for WT (r = 0.96) and WMH (r = 0.89). Larger lesion volumes were positively correlated with higher/better *Dice* scores for WT (r = 0.33), WMH (r = 0.34), and across all lesions (r = 0.89) on a log(10) transformed scale. While the median *Dice* for WMH was 0.42 across training subjects with WMH, the median *Dice* was 0.62 for those with

OPEN ACCESS

Edited by:
Markus Diesmann, Julich Research Centre, Germany

Reviewed by:
Qiang Luo, Fudan University, China
Sen Song, Tsinghua University, China

***Correspondence:**
Jeffrey D. Rudie
jeffrey.rudie@ucsf.edu

† These authors have contributed equally to this work and share senior authorship

Received: *28 August 2019*
Accepted: *04 December 2019*
Published: *20 December 2019*

Citation:
Rudie JD, Weiss DA, Saluja R, Rauschecker AM, Wang J, Sugrue L, Bakas S and Colby JB (2019) Multi-Disease Segmentation of Gliomas and White Matter Hyperintensities in the BraTS Data Using a 3D Convolutional Neural Network. Front. Comput. Neurosci. 13:84. doi: 10.3389/fncom.2019.00084

at least 5 cm^3 of WMH. We anticipate the development of computational algorithms that are able to model multiple diseases within a single subject will be a critical step toward translating and integrating artificial intelligence systems into the heterogeneous real-world clinical workflow.

Keywords: segmentation, glioblastoma, convolutional neural network, white matter hyperintensities, deep learning, radiology, multi-disease classification

INTRODUCTION

A significant challenge in the deployment of advanced computational methods into typical clinical workflows is the vast heterogeneity of disease processes, which are present both between individuals (inter-subject heterogeneity) and within individuals (intra-subject heterogeneity). Most adults over the age of 60 have a variable degree of abnormal signal on brain MRIs due to age-related changes manifesting as white matter hyperintensities (WMH), which are typically secondary to small vessel ischemic disease (SVID) and chronic infarcts that can be found in subjects with vascular risk factors and clinical histories of stroke and dementia (Wardlaw et al., 2015). These lesions can confound automated detection and segmentation of other disease processes, including brain tumors, which also result in abnormal signal in T2-weighted (T2) and T2 Fluid-attenuated inversion recovery (T2-FLAIR) MRI scans secondary to neoplastic processes and associated edema/inflammation. We sought to address this challenge of intra-individual heterogeneity by leveraging (i) the dataset of the International Multimodal Brain Tumor Segmentation (BraTS) 2018 challenge (Menze et al., 2015; Bakas et al., 2017b, 2019) (ii) expert radiologist expertise, and (iii) three-dimensional (3D) convolutional neural networks (CNNs).

Advances in the field of segmentation and radiomics within neuro-oncology have been supported by data made available through The Cancer Imaging Archive (TCIA; Clark et al., 2013). Since 2012, the BraTS challenge has further curated TCIA glioma multi-parametric MRI (mpMRI) scans, segmentation of tumor sub-regions, and survival data in a public dataset and sponsored a yearly challenge to improve performance of automated segmentation and prognostication methods (Menze et al., 2015; Bakas et al., 2017b, 2019). Similar to BraTS, there have been large efforts for improving automatic segmentation of WMH (Griffanti et al., 2016; Habes et al., 2016), which include the MICCAI 2017 WMH competition (Li et al., 2018; Kuijf et al., 2019), as well as stroke lesions, through the Ischemic Stroke Lesion Segmentation Challenge (ISLES; Winzeck et al., 2018). Deep learning (DL) approaches for biomedical image segmentation are now established as superior to the previous generation of atlas-based and hand-engineered feature approaches (Fletcher-Heath et al., 2001; Gooya et al., 2012), as demonstrated by their performance in recent image segmentation challenges (Chang, 2017; Kamnitsas et al., 2017; Li et al., 2018; Bakas et al., 2019; Myronenko, 2019).

Deep learning relies on hierarchically organized layers to process increasingly complex intermediate feature maps and utilizes the gradient of the error in predictions with regard to the units of each layer to update model weights, known as "back-propagation." In visual tasks, this allows for the identification of lower- and intermediate-level image information (feature maps) to maximize classification performance based on annotated datasets (LeCun et al., 2015; Chartrand et al., 2017; Hassabis et al., 2017). Typically, CNNs, a class of feed-forward neural networks, have been used for image-based problems, achieving super-human performance in the ImageNet challenge (Deng et al., 2009. Krizhevsky et al., 2012). The U-Net architecture (Ronneberger et al., 2015; Cicek et al., 2016; Milletari et al., 2016) describes a CNN with an encoding convolutional arm and corresponding decoding [de]convolutional arm has been shown to be particularly useful for 3D biomedical image segmentation through its semantic- and voxel-wise approach, such as for segmentation of abnormal T2-FLAIR signal across a range of diseases (Duong et al., 2019).

Several prior machine learning approaches have been used to model inter-subject disease heterogeneity, such as distinguishing on an individual subject basis between primary CNS lymphoma and glioblastoma (Wang et al., 2011), or between different types of brain metastases (Kniep et al., 2018). There is evidence that these approaches may be superior to human radiologists (Suh et al., 2018), yet little work has been done to address intra-subject lesion heterogeneity. Notably, one recent study used CNNs to distinguish between WMH due to SVID versus stroke, finding that training a CNN to explicitly distinguish between these diseases allowed for improved correlation between SVID burden and relevant clinical variables (Guerrero et al., 2018). Although a large body of work has detailed methodological approaches to improve segmentation methods for brain tumors, to the best of our knowledge no prior studies have addressed intra-subject disease heterogeneity in the BraTS dataset.

Although the task of distinguishing between different diseases within an individual is typically performed subconsciously by humans, distinguishing between different diseases could be challenging for an automated system if it were not specifically designed and trained to perform such a task. When provided with enough labeled training data, image-based machine learning methods have shown success in identifying patterns that are imperceptible to humans. These include GBM subtypes related to specific genetic mutations (i.e., radiogenomics; Bakas et al., 2017a; Korfiatis et al., 2017; Akbari et al., 2018; Chang et al., 2018; Rathore et al., 2018), or imaging subtypes that are predictive of clinical outcomes (Rathore et al., 2018). Therefore, we sought to train a 3D U-Net model to distinguish between abnormal radiographic signals arising from brain glioma versus WMH in individual subjects, in the mpMRI data of the BraTS 2018 challenge. We hypothesized that this would (1) allow

for automatic differentiation of different disease processes, and (2) improve overall accuracy of segmentation of brain tumor extent of disease, particularly in subjects with a large amount of abnormal signal due to WMH.

MATERIALS AND METHODS

Data

We utilized the publicly available data of the BraTS 2018 challenge that describe a multi-institutional collection of pre-operative mpMRI brain scans of 351 subjects ($n_{training}$ = 285, and $n_{validation}$ = 66) diagnosed with high-grade (glioblastoma) and lower-grade gliomas. The mpMRI scans comprise native T1-weighted (T1), post-contrast T1-weighted (T1PC), T2, and T2-FLAIR scans. Pre-processing of the provided images included re-orientation to LPS (left-posterior-superior) coordinate system, co-registration to the same T1 anatomic template (Rohlfing et al., 2010), resampling to isotropic 1 mm^3 voxel resolution and skull-stripping as detailed in Bakas et al., 2019. Manual expert segmentation of the BraTS dataset delineated three tumor sub-regions: (1) Necrotic core (NCR), (2) active tumor (enhancing tissue; AT), and (3) peritumoral edematous/infiltrated tissue (ED). The whole tumor extent (WT) was considered the union of all these three classes.

Manual Annotation of WMH

In order to define the new tissue class of abnormal signal relating to WMH in the BraTS training subjects, a neuroradiologist (JR; neuroradiology fellow with extensive segmentation experience) defined manually segmentation masks of WMH using ITK-SNAP (Yushkevich et al., 2006). WMH were considered to be abnormal signal due to SVID, chronic infarcts, and/or any periventricular abnormal signal contralateral to the tumor. Examples of these new two class segmentations of the BraTS 2018 dataset are shown in **Figure 1**.

U-Net Architecture

We adapted the 3D U-Net architecture (Cicek et al., 2016; Milletari et al., 2016) for voxelwise image segmentation. Our encoder-decoder type fully convolutional deep neural network consists of (1) an encoder limb (with successive blocks of convolution and downsampling encoding progressively deeper/higher-order spatial features), (2) a decoder limb (with a set of blocks – symmetric to those of the encoder limb – of upsampling and convolution, eventually mapping this encoded feature set back onto the input space), and (3) an introduced novel so-called skip connections (whereby outputs of encoding layers are concatenated with inputs to corresponding decoding layers) in order to improve spatial localization over previous generations of fully convolutional networks (3D Res-U-Net; Milletari et al., 2016).

Our adaptations from the prototypical U-Net architecture included: 4 channel input data (T1, T1PC, T2, T2-FLAIR), 4 or 5 class output data (background = 0, NCR = 1, ED = 2, AT = 4, WMH = 3), with 3D convolutions, and no voxelwise weighting of input label masks. Training patch size was 80 × 80 × 80 voxels (mm), and inference was conducted in the whole image. We zero padded the provided images to increase its size from 240 × 240 × 155 voxels to 240 × 240 × 160 voxels, and hence being divisible by the training input patch size (80 × 80 × 80). Training patch centerpoints were randomly sampled from within the lesion (90%) or from within the whole brain (10%). Train-time data augmentation was performed with random left-right flipping, and constrained affine warps (maximum rotation 45°, maximum scale ±25%, maximum shear ±0.1). Core convolutional blocks included two nodes each of 3D convolution (3 × 3 × 3 kernel, stride = 1, zero padded), rectified linear unit activation, and batch normalization. Four encoding/decoding levels were used, with 32 convolutional filters (channels) in the base/outermost level, and channel number increased by a factor of two at each level (**Figure 2**).

The network was trained on an NVIDIA Titan Xp GPU (12GB), using the Xavier initialization scheme, Adam optimization algorithm (Kingma and Lei Ba, 2015; initial learning rate $1e^{-4}$), and 2nd order polynomial learning rate decay over 600 epochs. Training time was approximately 4.5 h. 10-fold internal cross validation on the training set was used for hyperparameter optimization and intrinsic estimation

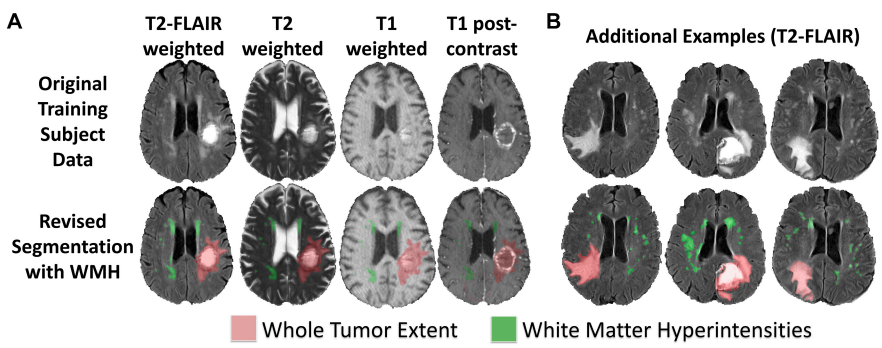


FIGURE 1 | Revised BraTS 2018 training segmentations including annotations of WMH. **(A)** Sample revised segmentation of abnormal signal due to WT (red) and WMH (green) on T2-FLAIR, T2, T1, and T1PC axial slices. **(B)** Three additional example revised segmentation maps for tumor and WMH overlaid on T2-FLAIR axial slices.

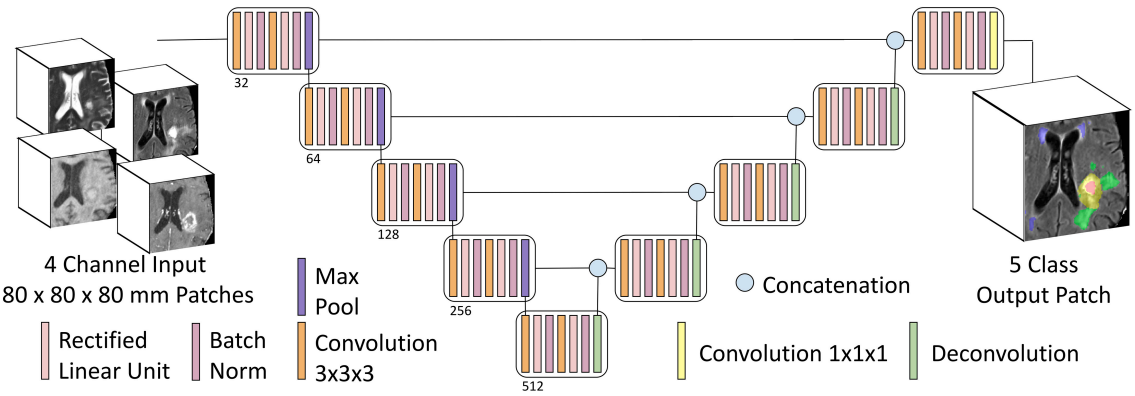


FIGURE 2 | Multiclass input and multiclass output U-Net schematic. Our U-Net has 4-channel input accepting 3D patches from mpMRI with four encoding and decoding layers, and either a four-class output (background, AT, NCR, and ED) or a five-class output (adding WMH as the fifth class).

of generalization performance during training. For inference on the validation set, the model was retrained 10 times independently on the *entire* training set (n = 285), and model predictions were averaged.

We trained models using this architecture twice; once with the four tissue classes originally annotated in the BraTS dataset, and again with the manual WMH segmentations added as a fifth class. All of the code has been made publicly available at https://github.com/johncolby/svid_paper.

Performance Metrics

Tissue segmentation performance was evaluated with the Dice metric ($2\times TP/(2\times TP + FP + FN)$; TP = true positive; FP = false positive; FN = false negative; Dice, 1945) for the tumor segmentation in both models, as well as for WMH in the five-class model. In addition, the 95th percentile of the Hausdorff distance (Hausdorff^{95}) was used as a performance evaluation metric, to evaluate the distance between the centers of the predicted and the expert 3D segmentations. The metrics for the four tissue classes and the Hausdorff^{95} distance were measured by submitting our segmentations to the online BraTS evaluation portal[1] (Davatzikos et al., 2018).

Further Exploration of U-Net Results

In order to further interrogate the performance of our proposed model, we performed correlations between manually segmented and predicted volumes for WT and WMH, as well as Bland Altman plots to assess agreement between the two measures of tissue volumes for both WT and WMH. For the evaluation of WMH, we performed correlations among the 196 cases that contained at least 100 mm^3 of WMH. To better understand what could affect performance, we also evaluated correlations between total lesion volumes and Dice scores.

[1]https://ipp.cbica.upenn.edu/

RESULTS

Manual WMH Segmentations

Of the manually revised BraTS training data (285 subjects), we found 196 (68.8%) with at least 100 mm^3 of WMH, 109 (38.4%) with at least 1000 mm^3 (1 cm^3) of WMH, 32 (11.2%) with at least 5000 mm^3 (5 cm^3) of WMH, and 17 (5.8%) with at least 10000 mm^3 (10 cm^3) of WMH. The manual WMH segmentations have been made available for public use at https://github.com/johncolby/svid_paper.

Segmentation Performance

The performance metrics for the training (10-fold cross validation) and validation subjects (final model) for each of the tissue classes in the four- and five-class models are shown in **Table 1**. We achieved a median Dice of 0.92 for WT in both the four- and five-class models, in both the training (p = 0.52; 10-fold cross validation) and validation datasets (p = 0.94). Segmentation performance on AT and tumor core (the union of AT and NCR) were also not significantly different between the four- and five-class models. There were no significant differences between tumor segmentation performance for high- or low-grade gliomas in the training set (p = 0.45).

The median Hausdorff^{95} distance in the training data was significantly lower (p = 0.002; two tailed t-test) in the five-class model (3.0, interquartile range 2.2–9.0) than the traditional four-class model (3.5, interquartile range 2.2–4.9). Example training cases where the Hausdorff^{95} distance were much better in the five-class model are shown in **Figure 3** with predicted segmentations for AT, NCR, ED, and WMH for both the four and five-class models. However, the Hausdorff^{95} distance was not significantly different in the validation data (p = 0.84). Example validation cases with greater than 5 cm^3 of WMH are shown in **Figure 4**, with predicted segmentations for AT, NCR, ED, and WMH.

We achieved a median Dice of 0.42 in the 189 subjects with WMH of at least 100 mm^3. Median Dice for WMH in subjects with at least 1000 mm^3 (1 cm^3), 5000 mm^3 (5 cm^3)

TABLE 1 | Performance metrics of four- and five-class models applied to the BraTS 2018 Training and Validation datasets.

	Training (n = 285)			Validation (n = 66)		
Performance metric	**4 Class model**	**5 Class model**	***p* value**	**4 Class model**	**5 Class Model**	***p* value**
Dice (Whole Tumor)	0.92 (0.87-0.95)	0.92 (0.87-0.94)	0.52	0.92 (0.89-0.95)	0.92 (0.90-0.95)	0.94
Dice (Enhancing Tumor)	0.82 (0.68-0.88)	0.82 (0.68-0.88)	0.76	0.87 (0.82-0.91)	0.87 (0.81-0.91)	0.99
Dice (Tumor Core)	0.88 (0.75-0.93)	0.89 (0.77-0.93)	0.75	0.91 (0.81-0.95)	0.91 (0.80-0.95)	0.97
Dice (WMH)	N/A	0.42 (0.25-0.55)	N/A	N/A	N/A	N/A
Hausdorff Distance (WT)	3.5 (2.2-9.0)	3.0 (2.2-4.9)	0.002	3.1 (2.0-4.5)	3 (2.0-4.4)	0.84

Dice scores are median (25th percentile – 75th percentile).

and 10000 mm^3 (10 cm^3) of WMH was 0.52, 0.62, and 0.67, respectively.

Correlation Between Predicted Lesion Volumes and Manual Segmented Volumes

Within the training dataset there was a strong correlation between manually segmented WT volume and predicted WT volume (Pearson r = 0.96, p < 0.0001; **Figure 5A**). There was also a strong correlation between manually segmented WMH volume and predicted WMH volume (Pearson r = 0.89, p < 0.0001; **Figure 5B**). Bland-Altman plots assessing agreement between manual and predicted volume for WT and WMH are shown in **Figures 5C,D**.

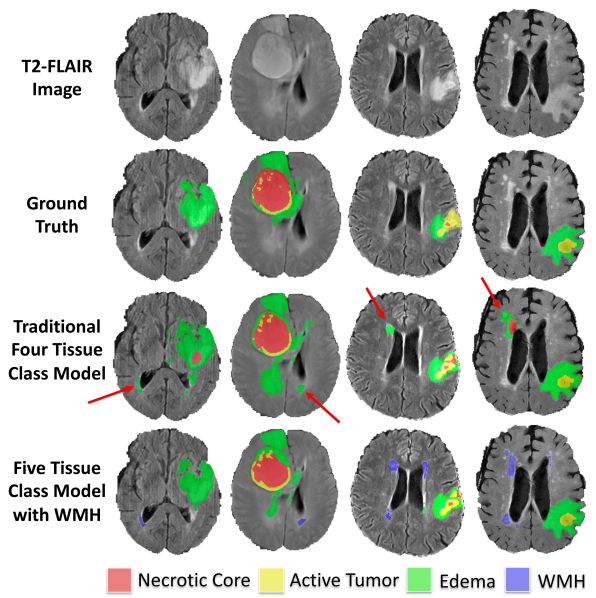


FIGURE 3 | Example segmentations in training subjects with smaller (better) Hausdorff distance metrics in the model with WMH. Axial T2-FLAIR slices of four example training cases are shown in the first row. The ground truth segmentations are overlaid in the second row (background, AT, NCR, and ED). The predicted tumor segmentations overlaid from the four-class model (background, AT, NCR, and ED) and the five-class model (background, AT, NCR, ED, and WMH) are shown in the third and fourth rows, respectively. The red arrows indicate multiple WMH distal to the tumor that were incorrectly classified in the four-class model as either ED or NCR.

Correlations Between Lesion Volumes and Dice

Within the training dataset there was a significant correlation between manually segmented WT volumes and WT Dice scores (Pearson r = 0.33, p < 0.0001; **Figure 6A**) and between manually segmented WMH volumes and WMH Dice scores (Pearson r = 0.34; p < 0.0001; **Figure 6B**). When combining WT and WMH, there was a stronger correlation between lesion volumes and Dice scores (Pearson r = 0.68; p < 0.0001; **Figure 6C**), which was even stronger when the volumes were transformed to a logarithmic (log(10)) scale (Pearson r = 0.89; p < 0.0001; **Figure 6D**). There was no significant relationship between WMH volume and WT Dice scores (Pearson r = −0.05, p = 0.42).

DISCUSSION

Advanced computational methods are poised to improve diagnostic and treatment methods for patients diagnosed with glioma (Davatzikos et al., 2019; Rudie et al., 2019). However, a critical challenge facing the eventual deployment of artificial

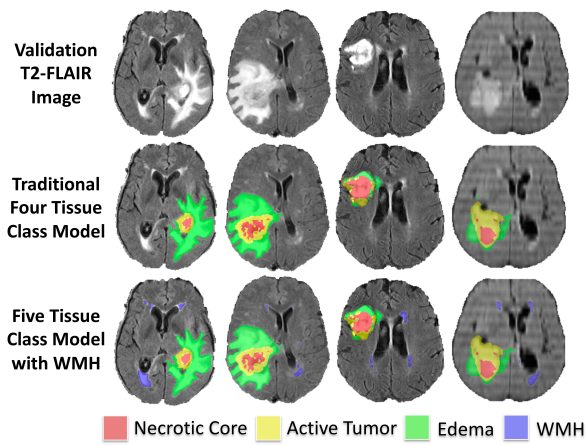


FIGURE 4 | Example predicted segmentations in validation subjects with greater than 5 cm^3 of WMH. Axial T2-FLAIR slices of four example validation cases are shown in the first row. The predicted tumor segmentations overlaid from the four-class model (background, AT, NCR, and ED) and the five-class model (background, AT, NCR, ED, and WMH) are shown in the second and third row, respectively.

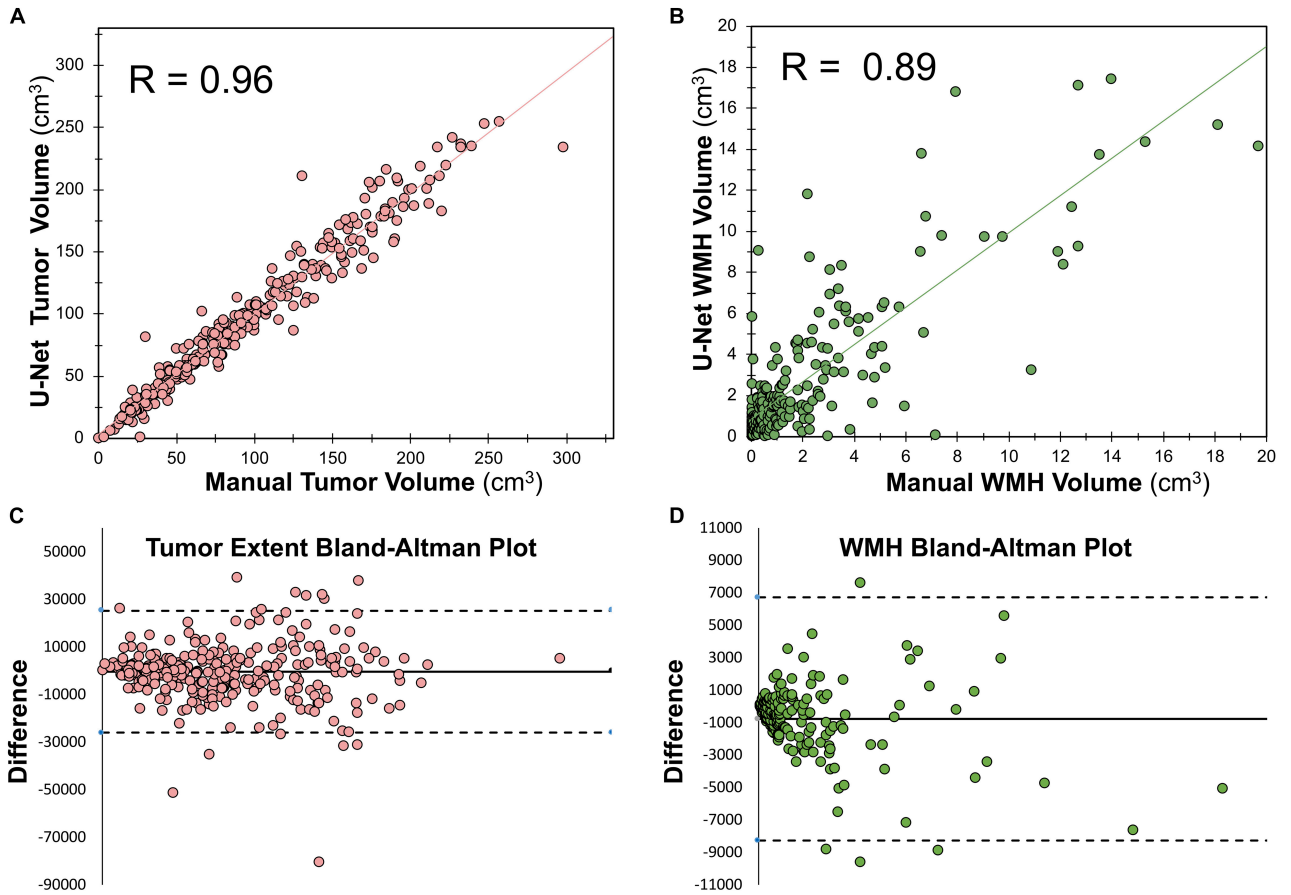


FIGURE 5 | Relationship between manually segmented volume and U-Net predicted volume. **(A)** Pearson correlation between manually segmented tumor volume and U-Net predicted tumor volume. **(B)** Spearman ranked correlation between manually segmented WMH volume and U-Net predicted WMH volume. **(C)** Bland–Altman plot for WT manually segmented volume and U-Net predicted volume. **(D)** Bland–Altman plot for WMH manually segmented volume and U-Net predicted volume. The dotted lines in panels **(C,D)** mark the bounds of the 95% confidence interval of the bias.

intelligence systems into daily clinical practice is disease heterogeneity within subjects. In this study, we utilized the BraTS 2018 dataset and expert-revised WMH segmentations to train a state-of-the-art CNN to successfully distinguish and quantify abnormal signal due to WMH as a distinct tissue class from glioma tissue sub-regions.

We used a 3D CNN (U-Net architecture; Cicek et al., 2016; Milletari et al., 2016) for multiclass tissue segmentation with performance at the top 10% of the BraTS 2018 leaderboard (Bakas et al., 2019; noting that we did not participate in the official competition). U-Nets have been particularly adept at medical image segmentation, due to their ability to convert feature maps obtained during convolutions into a vector and from that vector reconstruct a segmentation, which reduces distortion by preserving the structural integrity.

To our knowledge this is the first study to distinguish intra-subject lesion heterogeneity in the BraTS dataset, noting that Guerrero et al. (2018) previously used a U-Net architecture to distinguish chronic infarcts from WMH due to SVID. Although we hypothesized that adding WMH as a tissue class could improve tumor segmentation performance, we did not find a significant difference between tumor segmentation overlap (Dice) in the model that incorporated WMH as an additional class. Incorporating WMH as a distinct fifth-class did significantly (p = 0.002; two tailed t-test) improve the Hausdorff (95th percentile) distance metric within the training sample. As the Hausdorff^{95} distance reflects the center of the lesion, and WMH are often far from the tumor, poorer Hausdorff^{95} distance in the four-class model was likely due to false positive segmentations of WMH as tumor as demonstrated in **Figure 3**. However, upon reviewing validation cases with larger amounts of predicted WMH (**Figure 4**), it appeared that the original four-class model, although not explicitly trained to model WMH, mostly learned to implicitly ignore most WMH, likely due to spatial characteristics of the WMH being distant from the primary tumors and in characteristic locations and shapes. It is possible that the addition of WMH as an additional class could degrade segmentation performance of ED that was relatively distal to the center of tumor, thus the benefits of reducing distal WMH false positives in the five-class model may have be counterbalanced by increasing false negatives.

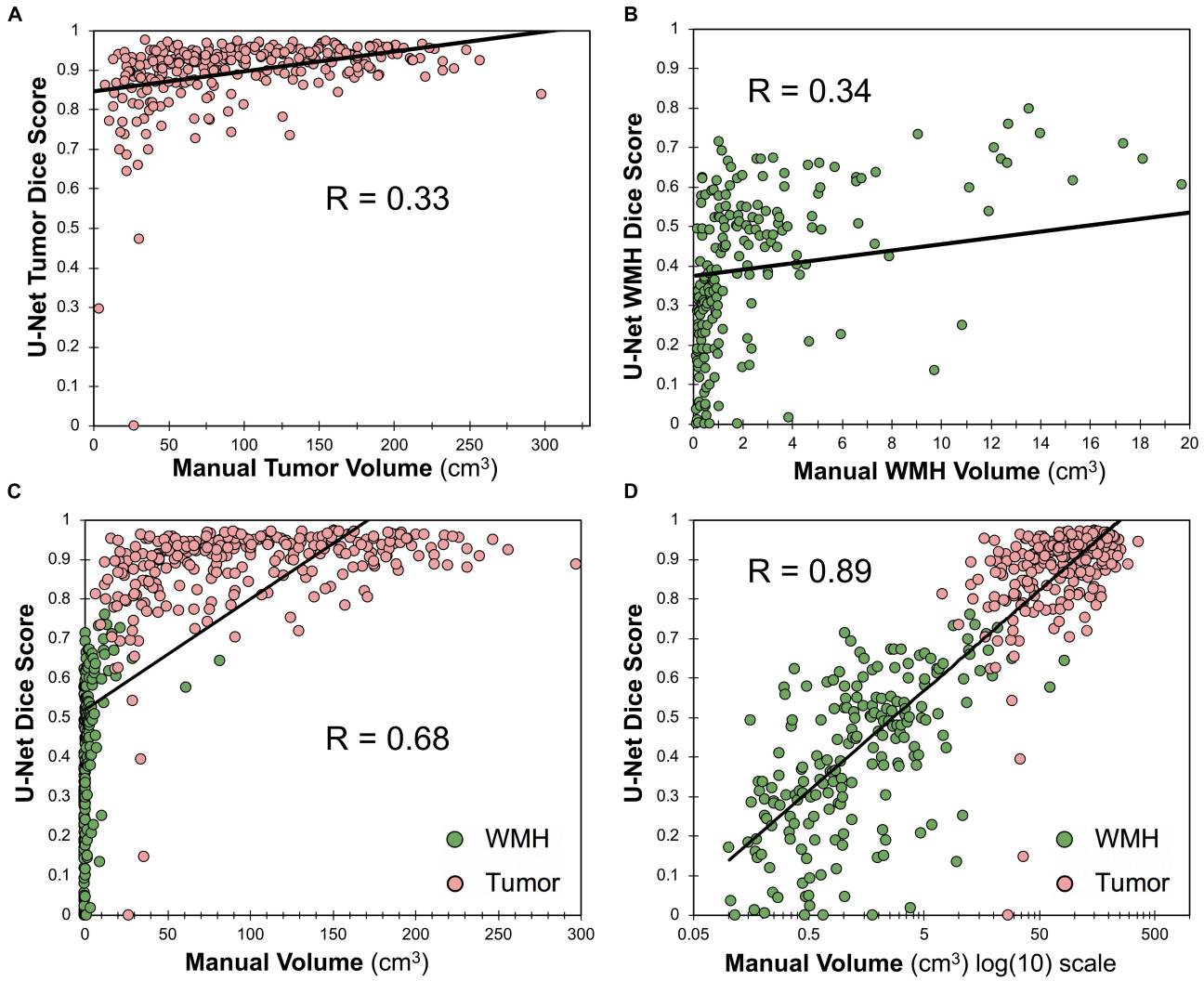


FIGURE 6 | Relationship between Dice score and lesion volume. **(A)** Scatter plot and Pearson correlation between WT volume and WT Dice score. **(B)** Scatter plot and Pearson correlation between WMH volumes and WMH Dice scores. **(C)** Scatter plot and Pearson correlation between volumes and Dice scores for both WMH and WT. **(D)** Scatter plot and Pearson correlation between log(10) transformed volumes and Dice scores for both WMH and WT.

As evidenced by the BraTS leaderboard, a Dice of ∼0.90 is considered excellent and has previously been shown to be at a level similar to inter-rater reliability for BraTS (Visser et al., 2019). As demonstrated in **Figure 6**, we found that lesion volume was an important predictor of Dice scores for both WT (**Figure 6A**) and WMH (**Figure 6B**). When evaluating both WT and WMH (**Figures 6C,D**), we found that the majority of the variance in Dice scores was explained by lesion volume, particularly when transformed to a logarithmic scale (**Figure 6D**). Thus, poorer performance for WMH in our data appear to largely be driven by smaller lesion sizes. This is consistent with prior literature that has also shown positive correlations between lesion volumes and Dice scores (Winzeck et al., 2018; Duong et al., 2019). Although our reported Dice scores for WMH appear relatively low (0.42), it should be noted that average volume of WMH in the MICCAI 2017 dataset was 16.9 cm^3 (Kuijf et al., 2019). When looking at cases with larger volumes of WMH (>10 cm^3) the average Dice score (0.67) was more similar to those reported in the 2017 MICCAI WMH dataset (0.70–0.80; Kuijf et al., 2019). A further explanation for reduced segmentation performance of smaller lesions may be lower inter-rater reliability, such as what has been reported in multiple sclerosis (Dice ∼0.60; Egger et al., 2017). A limitation of the current study is that there is only a single expert annotation for both the BraTS dataset and the WMH, thus the contribution of inter-rater reliability could not be assessed. In the future we also plan to improve detection of smaller lesions by using different neural network architectures, such as two-stage detectors (Girshick et al., 2014), or implementing different loss functions, such a focal loss (Lin et al., 2018; Abraham and Khan, 2019).

As artificial intelligence tools start to become integrated with clinical workflows for more precise quantitative assessments of

disease burden, it will be necessary to distinguish, quantify and longitudinally assess a variety of disease processes, in order to assist with more accurate and efficient clinical decision-making. Explicitly tackling intra-subject disease heterogeneity by training models to perform these tasks should help translate these advanced computational methods into clinical practice.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the BraTS 2018 dataset https://www.med.upenn.edu/sbia/brats2018/data.html. The manual WMH segmentations and code used in this manuscript have been made available for public use at https://github.com/johncolby/svid_paper.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JR and JC performed the analysis and prepared the initial manuscript. DW, RS, and JW performed some of the analyses. AR, LS, and SB helped to direct the research. All authors helped to revise the manuscript.

FUNDING

This study was partly supported by the institutional T-32 Training Grants from the NIH/NIBIB (Penn T32 EB004311-10 and UCSF T32-EB001631-14) and NIH/NINDS:R01NS042645 and NIH/NCI:U24CA189523 grants. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH. We would like to acknowledge the NVIDIA Corporation that donated three Titan Xp GPUs as part of the NVIDIA GPU grant program (awarded to JR, AR, and JC).

REFERENCES

Abraham, A., and Khan, K. (2019). "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *Proceedings of the IEEE Comput Soc Conf Comput Vis Pattern Recogn Proceedings - International Symposium on Biomedical Imaging 2019-April*, Piscataway, NJ.

Akbari, H., Bakas, S., Pisapia, J. M., Nasrallah, M. P., Rozycki, M., Martinez-Lage, M., et al. (2018). In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature. *Neuro Oncol.* 20, 1068–1079. doi: 10.1093/neuonc/noy033

Bakas, S., Akbari, H., Pisapia, J., Martinez-Lage, M., Rozycki, M., Rathore, S., et al. (2017a). *In Vivo* detection of EGFRvIII in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep peritumoral infiltration: the φ-index. *Clin. Cancer Res.* 23, 4724–4734. doi: 10.1158/1078-0432.CCR-16-1871

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Advancing The cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., and Rempfler, M. (2019). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1811.02629

Chang, P., Grinband, J., Weinberg, B. D., Bardis, M., Khy, M., Cadena, G., et al. (2018). Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am. J. Neuroradiol.* 39, 1201–1207. doi: 10.3174/ajnr.A5667

Chang, P. D. (2017). Fully convolutional deep residual neural networks for brain tumor segmentation. *Brain Lesion* 108–118.

Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C. J., et al. (2017). Deep learning: a primer for radiologists. *Radiographics* 37, 2113–2131. doi: 10.1148/rg.2017170077

Cicek, C., Abdulkadir, A., Brox, B., Ronneberger, R., and Lienkamp, L. (2016). "3D U-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, eds S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, W. Wells (Cham: Springer)

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

Davatzikos, C., Rathore, S., Bakas, S., Pati, S., Bergman, M., Kalarot, R., et al. (2018). Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J. Med. Imaging* 5:011018. doi: 10.1117/1.JMI.5.1.011018

Davatzikos, C., Sotiras, A., Fan, Y., Habes, M., Erus, G., Rathore, S., et al. (2019). Precision diagnostics based on machine learning-derived imaging signatures. *Magn. Reson. Imaging* 64, 49–61. doi: 10.1016/j.mri.2019.04.012

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409

Duong, M. T., Rudie, J. D., Wang, J., Xie, L., Mohan, S., Gee, J. C., et al. (2019). Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. *AJNR Am. J. Neuroradiol.* 40, 1282–1290. doi: 10.3174/ajnr.A6138

Egger, C., Opfer, R., Wang, C., Kepp, T., Sormani, M. P., Spies, L., et al. (2017). MRI FLAIR lesion segmentation in multiple sclerosis: does automated segmentation hold up with manual annotation? *Neuroimage Clin.* 13, 264–270. doi: 10.1016/j.nicl.2016.11.020

Fletcher-Heath, L. M., Hall, L. O., Goldgof, D. B., and Murtagh, F. R. (2001). Automatic segmentation of non-enhancing brain tumors in magnetic resonance images. *Artif. Intell. Med.* 21, 43–63. doi: 10.1016/s0933-3657(00)00073-7

Girshick, G., Donahue, D., Darrell, D., and Malik, M. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Piscataway, NJ.

Gooya, A., Pohl, K. M., Bilello, M., Cirillo, L., Biros, G., Melhem, E. R., et al. (2012). GLISTR: glioma image segmentation and registration. *IEEE Trans. Med. Imaging* 31, 1941–1954. doi: 10.1109/TMI.2012.2210558

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., et al. (2016). BIANCA (brain intensity AbNormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. *Neuroimage* 141, 191–205. doi: 10.1016/j.neuroimage.2016.07.018

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., et al. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using

convolutional neural networks. *Neuroimage Clin.* 17, 918–934. doi: 10.1016/j.nicl.2017.12.022

Habes, M., Erus, G., Toledo, J. B., Zhang, T., Bryan, N., Launer, L. J., et al. (2016). White matter hyperintensities and imaging patterns of brain ageing in the general population. *Brain* 139, 1164–1179. doi: 10.1093/brain/aww008

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2017). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, M. Reyes (Cham: Springer)

Kingma, D. P., and Lei Ba, J. (2015). "ADAM: a method for stochastic optimization," in *Proceedings of the Conference Paper at ICLR 2015*, Ithaca, NY.

Kniep, H. C., Madesta, F., Schneider, T., Hanning, U., Schönfeld, M. H., Schön, G., et al. (2018). Radiomics of brain MRI: utility in prediction of metastatic tumor type. *Radiology* 290, 479–487. doi: 10.1148/radiol.2018180946

Korfiatis, P., Kline, T. L., Lachance, D. H., Parney, I. F., Buckner, J. C., and Erickson, B. J. (2017). Residual deep convolutional neural network predicts MGMT methylation status. *J. Digit. Imaging* 30, 622–628. doi: 10.1007/s10278-017-0009-z

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).. ImageNet classification with deep convolutional neural networks

Kuijf, H. J., Biesbroek, J. M., de Bresser, J., Heinen, R., Andermatt, S., Bento, M., et al. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities; results of the WMH segmentation challenge. *IEEE Trans. Med. Imaging* 38, 2556–2568. doi: 10.1109/TMI.2019.2905770

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W., et al. (2018). Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *ArXiv* [Preprint],

Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). "Focal loss for dense object detection," in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, Piscataway, NJ.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Milletari, F., Navab, N., and Ahmado, S. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, Piscataway, NJ.

Myronenko, A. (2019). 3D brain mri tumor segmentation using autoencoder regularization. *Brainles* 11384, 311–320. doi: 10.1007/978-3-030-11726-9_28

Rathore, S., Akbari, H., Rozycki, M., Abdullah, K. G., Nasrallah, M. P., Binder, Z. A., et al. (2018). Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Sci. Rep.* 8:5087. doi: 10.1038/s41598-018-22739-2

Rohlfing, T., Zahr, N. M., Sullivan, E. V., and Pfefferbaum, A. (2010). The SRI24 multichannel atlas of normal adult human brain structure. *Hum. Brain Mapp.* 31, 798–819. doi: 10.1002/hbm.20906

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, eds N. Navab, J. Hornegger, W. Wells, A. Frangi (Cham: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Rudie, J. D., Rauschecker, A. M., Bryan, R. N., Davatzikos, C., and Mohan, S. (2019). Emerging applications of artificial intelligence in neuro-oncology. *Radiology* 290, 607–618. doi: 10.1148/radiol.2018181928

Suh, H. B., Choi, Y. S., Bae, S., Ahn, S. S., Chang, J. H., Kang, S. G., et al. (2018). Primary central nervous system lymphoma and atypical glioblastoma: differentiation using radiomics approach. *Eur. Radiol.* 28, 3832–3839. doi: 10.1007/s00330-018-5368-4

Visser, M., Müller, D. M. J., van Duijn, R. J. M., Smits, M., Verburg, N., Hendriks, E. J., et al. (2019). Inter-rater agreement in glioma segmentations on longitudinal MRI. *Neuroimage Clin.* 22:101727. doi: 10.1016/j.nicl.2019.101727

Wang, S., Kim, S., Chawla, S., Wolf, R. L., Knipp, D. E., Vossough, A., et al. (2011). Differentiation between glioblastomas, solitary brain metastases, and primary cerebral lymphomas using diffusion tensor and dynamic susceptibility contrast-enhanced MR imaging. *AJNR Am. J. Neuroradiol.* 32, 507–514. doi: 10.3174/ajnr.A2333

Wardlaw, J. M., Valdés Hernández, M. C., and Muñoz-Maniega, S. (2015). What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. *J. Am. Heart Assoc.* 4:001140. doi: 10.1161/JAHA.114.001140

Winzeck, S., Hakim, A., McKinley, R., Pinto, J. A., Alves, V., Silva, C., et al. (2018). ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. *Front. Neurol.* 9:679. doi: 10.3389/fneur.2018.00679

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015

Conflict of Interest: DW and RS are consultants for the company Galileo CDS and receive consultant fees for their work, which is not directly related to this manuscript.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 24 January 2020
doi: 10.3389/fncom.2020.00003

Novel Volumetric Sub-region Segmentation in Brain Tumors

Subhashis Banerjee[1,2]* *and Sushmita Mitra*[1]

[1] *Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India,* [2] *Department of CSE, University of Calcutta, Kolkata, India*

A novel deep learning based model called Multi-Planar Spatial Convolutional Neural Network (MPS-CNN) is proposed for effective, automated segmentation of different sub-regions *viz.* peritumoral edema (*ED*), necrotic core (*NCR*), enhancing and non-enhancing tumor core (*ET*/*NET*), from multi-modal MR images of the brain. An encoder-decoder type CNN model is designed for pixel-wise segmentation of the tumor along three anatomical planes (axial, sagittal, and coronal) at the slice level. These are then combined, by incorporating a consensus fusion strategy with a fully connected Conditional Random Field (CRF) based post-refinement, to produce the final volumetric segmentation of the tumor and its constituent sub-regions. Concepts, such as spatial-pooling and unpooling are used to preserve the spatial locations of the edge pixels, for reducing segmentation error around the boundaries. A new aggregated loss function is also developed for effectively handling data imbalance. The MPS-CNN is trained and validated on the recent Multimodal Brain Tumor Segmentation Challenge (BraTS) 2018 dataset. The Dice scores obtained for the validation set for whole tumor (*WT* : *NCR*/*NE* + *ET* + *ED*), tumor core (*TC* : *NCR*/*NET* + *ET*), and enhancing tumor (*ET*) are 0.90216, 0.87247, and 0.82445. The proposed MPS-CNN is found to perform the best (based on leaderboard scores) for *ET* and *TC* segmentation tasks, in terms of both the quantitative measures (viz. Dice and Hausdorff). In case of the *WT* segmentation it also achieved the second highest accuracy, with a score which was only 1% less than that of the best performing method.

Keywords: convolutional neural network, brain tumor segmentation, spatial-pooling and unpooling, conditional random field, multi-planar CNN, class imbalance

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Wenqi Li,
Nvidia, United States
Ujjwal Raghunandan Baid,
Shri Guru Gobind Singhji Institute of Engineering and Technology, India
Raghav Mehta,
McGill University, Canada

***Correspondence:**
Subhashis Banerjee
mail.sb88@gmail.com

Received: *16 July 2019*
Accepted: *08 January 2020*
Published: *24 January 2020*

Citation:
Banerjee S and Mitra S (2020) Novel Volumetric Sub-region Segmentation in Brain Tumors. Front. Comput. Neurosci. 14:3. doi: 10.3389/fncom.2020.00003

1. INTRODUCTION

Gliomas (tumors of glial cells) represent 40% of tumors of the Central Nervous System, and 80% of all malignant brain tumors. The World Health Organization (WHO) grades these tumors based on the aggressiveness and infiltrative nature of their cells. Low-grade gliomas (LGG) are categorized as lowest- and intermediate-grades (WHO grades II and III), while high-grade gliomas (HGG) or glioblastoma constitute the highest-grade (WHO grade IV) (Louis et al., 2016). Diffuse LGGs are infiltrative brain neoplasms which affect different histological classes, and are called astrocytomas, oligodendrogliomas, and oligoastrocytomas (Louis et al., 2016). Although LGG patients are observed to have better survival than those with HGG, they often progress to secondary glioblastomas (GBMs) and eventual death (Li et al., 2013).

Accurate detection of tumor regions makes the job of the medical practitioner simpler, by allowing (i) appropriate measurement of tumor volume, (ii) growth monitoring of tumor in patients over time, and (iii) prognosis, with follow-up evaluation, and prediction of overall survival (OS). Based on the histological heterogeneity observed within a glioma tumor, its cells are partitioned into different sub-regions, i.e., peritumoral edema (ED), necrotic core (NCR), enhancing and non-enhancing tumor core (ET / NET) (Menze et al., 2015; Bakas et al., 2018). These sub-regions reflect important and clinically relevant information.

Magnetic Resonance Imaging (MRI) has become the standard non-invasive technique for brain tumor diagnosis, over the last few decades, due to its inherent improved soft tissue contrast (DeAngelis, 2001; Cha, 2006). MR imaging can effectively capture the intrinsic heterogeneity of gliomas using multimodal scans with varying intensity profiles. Typically four MR sequences *viz.* native $T1$-weighted ($T1$), $T2$-weighted ($T2$), post-contrast enhanced $T1$-weighted ($T1C$), and $T2$-weighted with FLuid-Attenuated Inversion Recovery (FLAIR), are used. The rationale behind using multiple sequences is the fact that different tumor regions are properly visible in different sequences, which again are complementary to each other; thereby rendering them as effective tools for accurately demarcating and distinguishing between different types of tumors (Banerjee et al., 2016a, 2017). Since gliomas are infiltrative, the sub-regions appear highly heterogeneous in MRI scans. Therefore, segmentation of Glioma sub-regions is considered to be one of the most challenging tasks in medical image analysis (Bakas et al., 2018).

Although manual segmentation of tumors is considered as the gold standard, it is time-consuming and prone to errors due to human fatigue. Therefore, there is a growing body of literature on computational algorithms, addressing this important task through supervised and unsupervised techniques (Menze et al., 2015; Banerjee et al., 2016b, 2018a,b; Mitra et al., 2017; Bakas et al., 2018). Development of such computer-aided tumor segmentation algorithms entails a lot of challenges due to the large spatial and structural variability among brain tumors. For example, segmenting HGG and LGG tumors with the same algorithm is a difficult proposition. It is also hard to compare any segmentation method with other existing ones, since they were often designed and validated on different private datasets. Such difficulty is due to various critical factors like (i) modalities used for the segmentation, (ii) state of the disease in which the image was taken (prior to treatment, or post-operative), (iii) type of the tumor (GBM or LGG, solid or infiltratively growing, primary or secondary), and can significantly influence the segmentation results.

Studies on tumor segmentation from brain MR images have been abundant in the literature. Here we provide a very recent literature review of the field. For extensive review on prior techniques, the reader is referred to (Bauer et al., 2013; Gordillo et al., 2013). Methodologically segmentation of tumors from brain MRI images can be broadly categorized under *generative* (Cuadra et al., 2004; Zacharaki et al., 2008; Menze et al., 2010; Banerjee et al., 2018a) and *discriminative* (Bauer et al., 2011; Zikic et al., 2012a,b; Wu et al., 2014; Menze et al., 2015; Bakas et al., 2018) family of models.

Generative methods are explicitly designed according to the anatomy and appearance of the tumor and the brain, and incorporate *a-priori* information for decision-making. Tumors can be modeled as outliers as compared to the expected shape and anatomy of the brain, as reported in references (Cuadra et al., 2004; Zacharaki et al., 2008). Menze et al. designed a generative probabilistic model for channel-specific segmentation of the tumor MRI in Menze et al. (2010). The generative approach in references (Gooya et al., 2012) first computes the spatial a-priori or "atlas" from healthy brain MRI scans. This is next modified using an expectation maximization (EM) algorithm, over a given set of patient images, to detect the most likely localization of the tumor therein. The concept of visual saliency is used in references (Banerjee et al., 2016b, 2018a; Mitra et al., 2017) for identifying tumor regions from brain MR images. This helps in automatically and quickly isolating the tumor region to be subsequently used for delineation. However, generative models are found to not generalize appropriately on unseen data; mainly due to their simple hypothesis functions. Their dependence on *a-priori* knowledge also makes them unsuitable to applications where this is not available.

On the other hand, *discriminative models* directly learn patterns from representation in the form of image features from the underlying training data, while not depending on any *a-priori* knowledge. These models may overfit the underlying training data, but have been shown to consistently perform well over unseen data due to their complex learned hypotheses. A hierarchical fully automated approach was presented (Bauer et al., 2011) for brain tissue segmentation, using support vector machine and conditional random fields. A combination of discriminative and generative models were developed (Zikic et al., 2012a) for the segmentation of high grade gliomas into the constituent sub-regions. This approach used decision forest as the discriminative classifier, which was fed with three unique, parameterized, contextually, and spatially aware features along with probabilities generated from Gaussian mixture models (Zikic et al., 2012b). Initial probability estimates were then used with spatially non-local features and context-sensitive decision forest for the classification of each data point. Another discriminative approach (Wu et al., 2014) used superpixels extracted from multi-modal MR images, with an SVM classifier being trained with features extracted by Gabor wavelet filters. A model-aware affinity model was defined, with its output being used alongside the SVM for application of conditional random fields theory before tumor segmentation.

Recently, Convolutional Neural Networks (LeCun et al., 1998) (CNNs or ConvNets) have been shown to work impressively on image recognition or classification problems (Krizhevsky et al., 2012). ConvNets are particularly useful for data that comes in the form of multiple arrays, like a color image. ConvNets essentially revolutionized the field of computer vision and have since become the de-facto standard for various object detection and recognition tasks (Farabet et al., 2013; Goodfellow et al., 2013; Sermanet et al., 2013; Simonyan and Zisserman, 2014). Inspired by their success, several medical imaging researchers have

applied them toward abnormality detection and segmentation; particularly, for brain MRIs. 3D ConvNets were used as a voxel wise classifier (Urban et al., 2014). Instead of looking at each slice of each sequence, the 3D ConvNet works directly with the volumetric MRI sequences; classifying each voxel into tumor or background. The problems with this approach are the high computational cost incurred during training and testing phases, as well as the requirement of huge datasets. A similar approach was used (Zikic et al., 2014) with minimal pre-processing, by looking at the 3D patch around each point in the sequence and classifying the central point as one of the labels. A two-way ConvNet architecture was developed (Havaei et al., 2017) to exploit both local and global contexts of the input image. Each pixel in every 2D slice of the MRI data was classified into one of the four tumor sub-regions or background, by predicting the label of the center pixel of an $M \times M$ patch. The idea of local structure prediction was transferred (Havaei et al., 2017) to the task of predicting dense labels of pathological structures in multi-modal 3D volumes using patch-based label dictionaries. Two separate ConvNet architectures were designed (Pereira et al., 2016) for HGG and LGG-pixel wise label prediction, along with the use of small kernels of size 3×3 throughout the ConvNets. An ensemble of ConvNet architectures (Kamnitsas et al., 2018) was introduced for robust brain tumor segmentation. The contribution won the multimodal brain tumor segmentation challenge (BraTS) in 2017. Three popular ConvNets, such as "DeepMedic" (Kamnitsas et al., 2016), "Fully Convolutional Network (FCN)" (Long et al., 2015), and "U-Net" (Ronneberger et al., 2015) were used to generate the class-confidence of each voxel in a multimodal MRI volume, with a class having the highest confidence being assigned to be the segmentation label of that voxel.

Inspired by the success of ConvNets in brain tumor segmentation, we propose here a new deep learning method for segmentation of different sub-regions viz. ED, NCR, ET, and NET, from multi-modal MR images of the brain. An encoder-decoder type ConvNet model is designed for pixel-wise segmentation of the tumor along three anatomical planes (axial, sagittal, and coronal) at the slice level. These are then combined, using a consensus fusion strategy with a fully connected Conditional Random Field (CRF) based post-refinement (Krähenbühl and Koltun, 2011), to produce the final volumetric segmentation of the tumor and its constituent sub-regions. Novel concepts, such as spatial-pooling and unpooling (Badrinarayanan et al., 2017) are used to preserve the spatial locations of the edge pixels, for reducing segmentation error around the boundaries. A new aggregated loss function is also developed for effectively handling data imbalance.

The rest of the paper is organized as follows. Section 2 describes details of data, preparation of the patch database for ConvNet training, the proposed multi-planar Spatial-ConvNet model which uses a spatial-pooling layer, the aggregated loss function for imbalanced data handling during segmentation, and the radiomic analysis of the segmented volume of interest for overall survival prediction. Section 3 provides experimental results on the segmentation in multi-planar and multi-sequence data, with overall survival prediction. It also demonstrates their effectiveness through qualitative and quantitative analysis. Finally section 4 draws conclusions, and provides directions for future research.

2. MATERIALS AND METHODS

In this section we present a detailed description of the brain tumor MRI dataset, and the proposed methods for tumor segmentation and patient overall survival (OS) prediction. Segmentation comprises of extraction of patches, training and testing of the segmentation model, and post-processing. The OS prediction consists of quantitative feature extraction and dimensionality reduction.

2.1. Dataset

Multi-modal MRI volumes used in this paper, were taken from the Multimodal Brain Tumor Segmentation Challenge (BraTS) 2018[1] (Menze et al., 2015; Bakas et al., 2017a,b,c, 2018). The dataset consists of 210 HGG and 75 LGG glioma cases as training, with 66 unlabeled (HGG or LGG) cases as validation samples. Multi-modal or multi-channel MRI volumes, consisting of $T1$, $T1C$, $T2$, and $FLAIR$, are available for each patient with the MRI volume being composed of 155 slices of 240×240 resolution. The MRI volumes are first carefully aligned to the same anatomical template, skull-stripped, and interpolated to 1mm^3 voxel resolution, before being made available for experimentation. Manual segmentation of the tumor sub-regions is done by experts, following the same annotation protocol for all patients. Their annotations were revised and approved by board-certified neuro-radiologists. Finally, the predicted labels are evaluated by merging three regions viz. whole tumor ($WT : NCR/NE + ET + ED$), tumor core ($TC : NCR/NET + ET$), and enhancing tumor (ET) as shown in **Figure 1**.

2.2. ConvNet for Tumor Segmentation

Here we present the proposed multi planar ConvNet architecture for automatic segmentation of different tumor sub-regions, i.e., ED, ET, and NCR/NET from a given multi-modal MRI scan. Novel spatial max pooling and unpooling layers are introduced to better approximate the tumor anatomical structure by minimizing segmentation errors around the tumor boundary during up sampling. An adaptive fusion strategy for accurate and robust segmentation, by combining output from the three principal planes (axial, coronal, and sagittal), is described. A weighted aggregated loss function is introduced to train the networks in the presence of class imbalance.

2.2.1. Patch Based Learning

Tumors are typically heterogeneous, depending on cancer subtypes, and contain a mixture of structural and patch-level variability. Applying a ConvNet directly to the entire slice has its inherent drawbacks. Since the size of each slice is 240×240, therefore overall memory requirement of the model will increase. Moreover, very little difference is observed in adjacent MRI slices at the global level; whereas, patches generated from the same slice often exhibit significant dissimilarity. We develop a Fully Convolutional Network (FCN) architecture for pixel-wise

[1] https://www.med.upenn.edu/sbia/brats2018/data.html

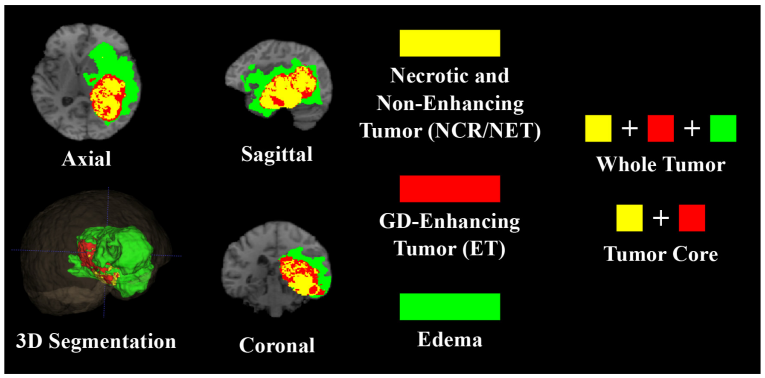


FIGURE 1 | T1 MRI of a sample HGG patient with 3D segmentation of different intra-tumoral structures (ED, ET, and NCR/NET) along three principal planes (axial, sagittal, and coronal).

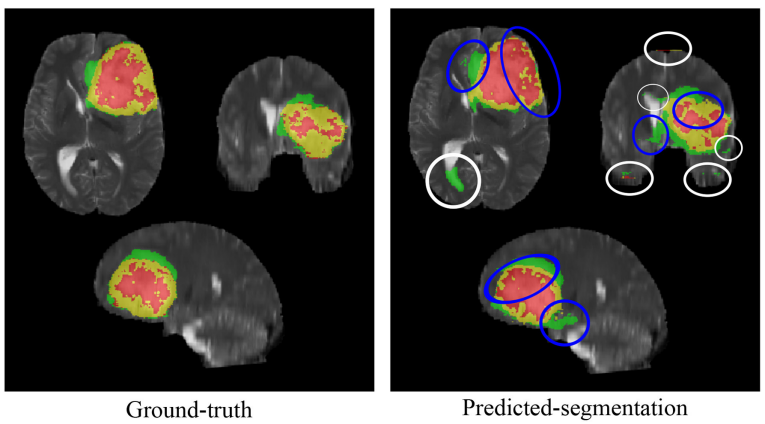


FIGURE 2 | Segmentation errors, with error around the boundary marked by blue ellipse and false positive errors are marked by white ellipses.

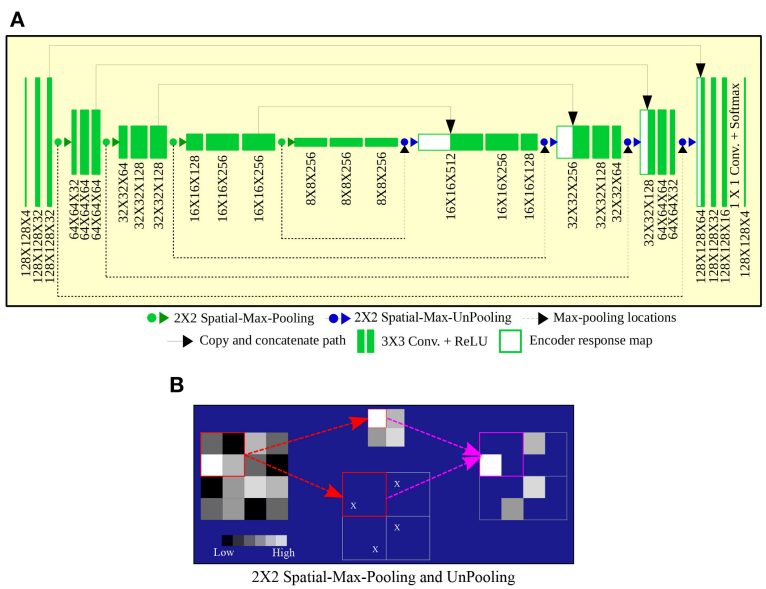


FIGURE 3 | (A) ConvNet architecture, with **(B)** Spatial-Max-Pooling and Unpooling, for segmentation.

segmentation of the tumor regions. Since FCN does not contain fully connected layers, it is invariant to input image size. Therefore, we can use images of different resolutions during training and testing (or inference).

2.2.2. ConvNet Architecture

The FCN architecture consists of three blocks "encoder or downsampling path," "bottleneck," and "decoder or upsampling path." The encoder block contains four feature extraction blocks, each having two consecutive convolution layers with filter (or kernel) size 3×3. Four max-pooling layers of window size 2×2 are placed in between the feature extraction blocks, to down sample an image into a set of high-level features. Pairs of convolution layers are placed in the bottleneck block, between the encoder and decoder blocks. The structure of decoder block is the same as that of the encoder, with the only difference being in the use of upsampling layer instead of max-pooling to construct a pixel-wise segmentation of the input MR patch.

It was observed during model validation that the predicted segmentation suffers mainly from two types of errors, as shown in **Figure 2**; (i) error around the boundary, and (ii) false positive at the top and bottom ends of the MRI volume. The error around the boundary occurs because the network loses spatial information during down sampling or pooling operations. The unpooling layers in the decoder block try to approximate the inverse of the pooling operation or upsample the reduced image to its original resolution through interpolation. In this process, the segmentation error percolates around the boundary of the region-of-interest (ROI) or volume-of-interest (VOI). This is considered as an important concern for a good medical image segmentation method. We name this as error around the boundary. The false positives error occur because the model is trained on 2D MRI patches without considering volumetric information.

2.2.3. Spatial-Max-Pooling and Unpooling

To circumvent the problem of error around the boundary to some extent, we used a modified version of the pooling and unpooling layers as proposed in references (Badrinarayanan et al., 2017)—and call it "spatial-max-pooling" and "spatial-max-unpooling." Now spatial-max-pooling can retain the position from where the max-pooling operation selected the maximum value, to be subsequently used during unpooling through the spatial-max-unpooling layer. Details of the process is illustrated in **Figure 3B**. Although the spatial-max-pooling and unpooling layers offer an advantage over regular nearest neighbor upsampling or deconvolution, they also increase the memory requirement of the overall model. Therefore, the max pooling locations for each of the input activation maps need to be stored for a mini batch, during each such operation, and reused in subsequent mini batches. Shortcut connections are used to copy and concatenate the high resolution response maps from the encoder to the decoder. It helps the decoder network localize and recover the object details more effectively. In this way we achieve a perfect agreement between high level features and pixel level details. **Figure 3A** illustrates the complete architecture of the proposed ConvNet model.

TABLE 1 | Hyperparameters used for training.

Model	Hyperparameters	Value	
CNN	Weights and bias	Xavier (Glorot and Bengio, 2010)	
	Optimizer	ADAM (Kingma and Ba, 2014)	
	Epochs	25	
	Batch_size	16	
	Learning rate	$1e^{-}4$	
	Hyperparameters	**Selected values**	**Values searched**
CRF	ω_1, ω_2	2.5, 4.0	[2, 2.5, 3, 3.5, 4], [2, 2.5, 3, 3.5, 4]
	$\sigma_{\alpha,x}$, $\sigma_{\alpha,y}$, $\sigma_{\alpha,y}$	24, 24, 24	[12, 24], [12, 24], [12, 24]
	$\sigma_{\beta,x}$, $\sigma_{\beta,y}$, $\sigma_{\beta,z}$	17, 12, 10	[10 − 20], [10 − 20], [6, 8, 10, 12]
	$\sigma_{\gamma,c}$	8	[4, 8, 12, 16]

2.2.4. Multi-Planar Aggregation With 3D CRF Based Refinement

The MRI scans are taken in the axial (X-Z) plane, which represents voxels (or an unit volume) of the 3-Dimensional human brain. Therefore, it can be reconstructed into coronal (Y-X) plane and sagittal (Y-Z) planes for having different 3D views of the brain. Using the multi-view property of MR imaging, we propose a solution for the second error, i.e., false positive error. We train three separate ConvNets (same architecture as **Figure 3A**) for segmenting the tumor along the three individual planes/views. Next the predicted probability maps generated by the softmax layers of the three ConvNets (p_{axial}, $p_{coronal}$, $p_{sagital}$) are fused by averaging the probability maps, i.e., $p = (p_{axial} + p_{coronal} + p_{sagittal})/3$. It is found that the integrated prediction from multiple planes are superior as compared to the estimated region based on any single plane in terms of accuracy, and robustness of decision. This is due to utilization of more information and minimization of the estimated loss.

Next a 3D fully-connected Conditional Random Field (CRF) based bilateral filtering (Krähenbühl and Koltun, 2011) is used to refine the fused prediction, while maintaining the local and contextual consistency of the segmentation. The 3D CRF integrates the four MRI sequences with the multi-planar fused predicted probability map, to produce an optimized segmentation by minimizing the energy function

$$E = \sum_i -\log p_i^{(l)} + \zeta(l_i, l_j)[\omega_1 \mathcal{P}(\lambda_i, \lambda_j) + \omega_2 f(\lambda_i, \lambda_j)], \quad (1)$$

where

$$\mathcal{P}(\lambda_i, \lambda_j) = \exp\left(-\sum_{d\in\{x,y,z\}} \frac{|s_{i,d} - s_{j,d}|}{2\sigma_{\alpha,d}^2}\right), \quad (2)$$

$$f(\lambda_i, \lambda_j) = \exp\left(-\sum_{c\in\{T1,T1C,T2,FLAIR\}} \frac{|I_{i,c} - I_{j,c}|}{2\sigma_{\gamma,c}^2} - \sum_{d\in\{x,y,z\}} \frac{|s_{i,d} - s_{j,d}|}{2\sigma_{\beta,d}^2}\right). \quad (3)$$

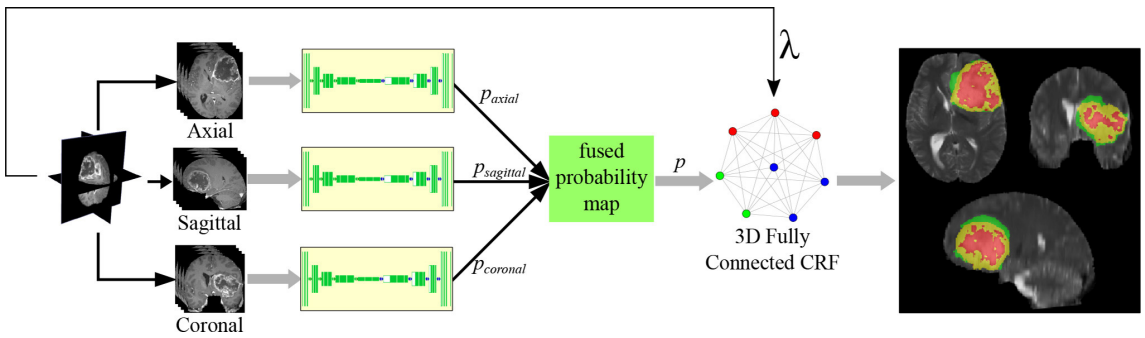


FIGURE 4 | Aggregated architecture combining multiple planes, with CRF-based refinement.

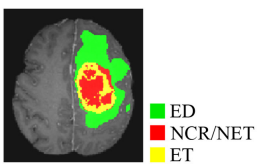
Here $p_i^{(l)}$ is the fused probability of assigning label l to voxel i and $\zeta(l_i, l_j)$ is the label compatibility function between voxel pairs $[l_i \neq l_j]$, with λ_i being the feature vector of voxel i containing seven features (viz. four intensities from the four MR sequences along with its 3D coordinate values). Note that $I_{i,c}$ corresponds to the intensity of the ith voxel in the four MRI sequences denoted by c, and $s_{i,d}$ represents the spatial 3D location of the voxel i. While function $\mathcal{P}(\cdot)$ controls the smoothness of the segmented region by considering the influence of neighborhood (using the hyperparameter $\sigma_{\alpha,d}$), the function $f(.)$ strives to preserve local and contextual consistency of the segmented output by controlling the level of similarity and proximity (using hyperparameters $\sigma_{\gamma,c}$ and $\sigma_{\beta,d}$). Optimizing the energy function also removes small isolated regions from the segmented output. All the model hyperparameters (α_1, α_2, σ_α, σ_γ, σ_β) are chosen through grid searching, as reported in **Table 1**.

The final model, represented in **Figure 4**, includes spatial-max-pooling and unpooling, multi-planar aggregation and 3D fully connected CRF based refinement. This will be referred to as "MPS-CNN" in the sequel.

2.2.5. Loss Function for Handling Class Imbalance

Since the dataset is highly imbalanced, with around 98% of the voxels belonging to either the healthy tissue or to the black surrounding area (as depicted in **Figure 5**), standard loss functions used in the literature are not suitable for training and optimizing the ConvNet. In such cases training can be dominated by the most prevalent class, with the classifiers focusing on learning the larger classes; thereby resulting in poor classification accuracy for the smaller classes. Therefore, we propose a new loss function. It is a sum of two factors viz.—Weighted Generalized Dice Loss (*WGDL*) (Sudre et al., 2017) and Weighted Log Loss (*WLL*) (Ronneberger et al., 2015). Both loss functions are computed between the soft binary segmentation or the probability map generated by the network using the softmax layer (P), and the corresponding gold standard/ground-truth image (G). The *WGDL* and *WLL* are defined as

$$WGDL = 1 - 2\frac{\sum_{c=1}^{|C|} wa_c \sum_{n=1}^{N} G_{cn}P_{cn}}{\sum_{c=1}^{|C|} wa_c \sum_{n=1}^{N} G_{cn} + P_{cn}}, \quad (4)$$



Class	Voxel Count	%
Background	8790699	98.46
ED	90806	1.02
ET	26264	0.29
NCR/NET	20231	0.23

FIGURE 5 | Tumor sub-class distribution for a sample MRI slice.

and

$$WLL = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{|C|} ws_c G_{cn}\log(P_{cn}), \quad (5)$$

where $C = \{Background, ED, ET, NCR/NET\}$, N is the total number of pixels in the image. Here the contribution of each class is multiplied by the adaptive weight $wa_c = \frac{1}{(\sum_{n=1}^{N} G_{cn})^2}$, which is inversely proportional to the class volume. Thereby it controls the contribution of larger classes while helping to learn smaller classes by reducing the classifier bias. Here ws_c is a four dimensional vector, storing the static class weights for $[Background, ED, ET, NCR/NET]$, and is assigned based on the class ratio. Parameters G_{cn} and P_{cn} correspond to the ground truth value and the predicted output, respectively, for the nth pixel w.r.t. the cth class. Optimizing the Generalized Dice Loss (*WGDL*) produces over segmented regions, while log loss generates under-segmented regions. Therefore, we combine *WGDL* and *WLL* in a weighted fashion, so that while cross-entropy treats every pixel as an independent prediction, the dice-score looks at the resulting mask in a more holistic manner. Moreover, considering the fact that these two losses yield significantly different masks, each with its own merits and errors, a combination of such complementary information should be beneficial.

3. EXPERIMENTAL SETUP AND RESULTS

The ConvNet models were developed using TensorFlow, with Keras in Python. The experiments were performed on the Intel AI DevCloud platform having cluster of Intel Xeon Scalable

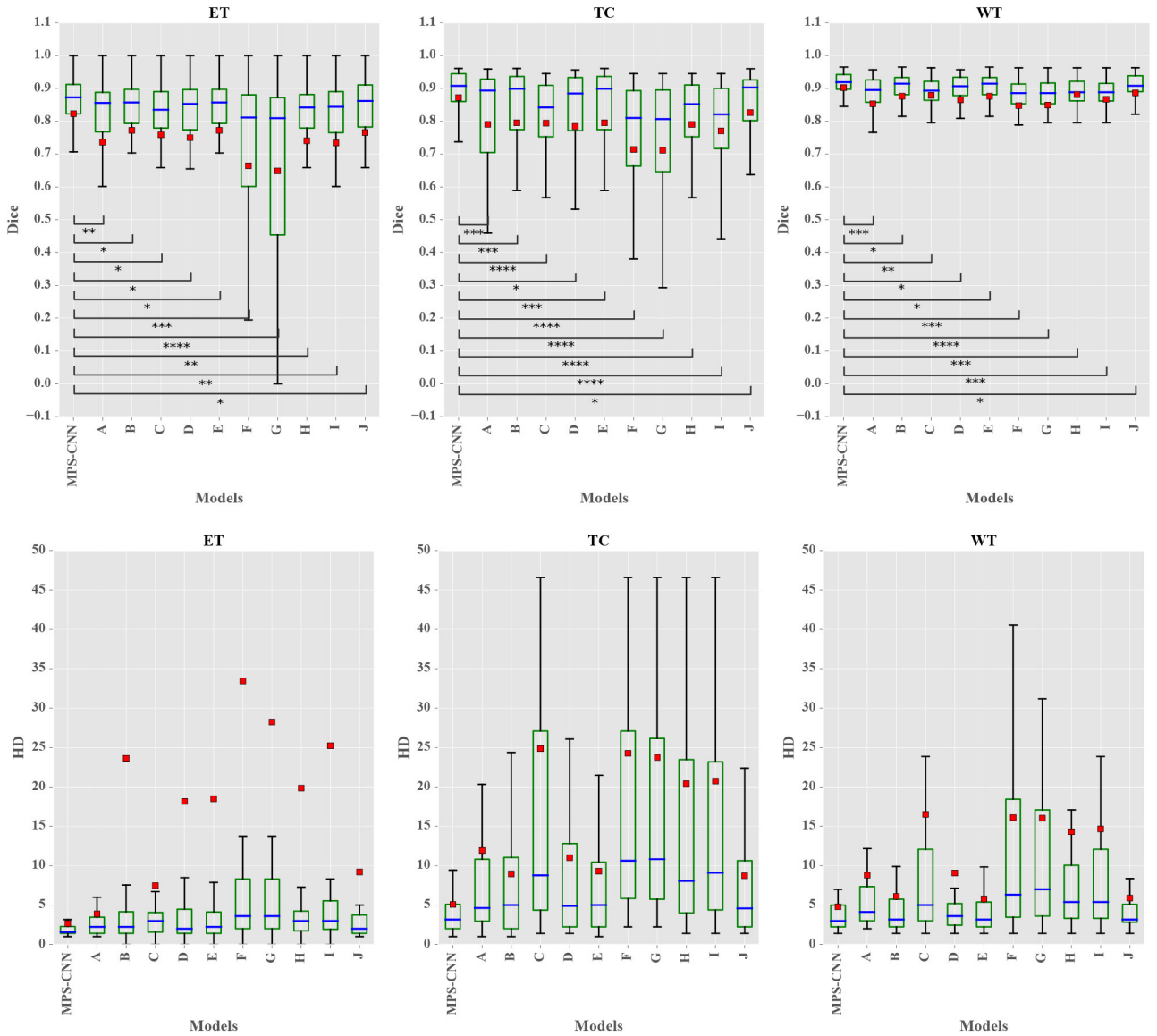


FIGURE 6 | Box plots of segmentation performance for the proposed MPS-CNN and the other 10 (A–J) models, measured by Dice score and Hausdorff distance, for the WT, TC, and ET tumor sub-regions of 66 patients from the BraTS 2018 validation dataset. The *p*-values <0.05, <0.001, <0.0001, and <0.00001 for each comparison are represented by *, **, ***, and ****, respectively, w.r.t. MPS-CNN.

processors and 96 GB of RAM. The proposed segmentation model was trained and validated on the corresponding training and validation datasets provided by the BraTS 2018 (Menze et al., 2015; Bakas et al., 2017a,b,c, 2018) organizers and is described in section 2.

The CNN models were trained on the patches extracted from the standardized and cropped MRI volumes. The BraTS 2018 datasets contains MRI volumes of size 155 × 240 × 240, which are cropped to have a size of 146 × 192 × 152 for discarding some unwanted background. This helps minimize the number of patches extracted from the "non-brain" region. Then patches of size 128 × 128 (experimentally found to be the best) were extracted randomly from all the four MRI sequences, with a constraints such that the center pixel of a patch does not belong to the minimum intensity value in the *FLAIR* modality. This condition helps minimize the extraction of "non-tumor" patches. A total of 111,690, 142,160, 118,400 training patches were extracted from the axial, coronal and sagittal planes, respectively. During inference the entire stack of slices (155 × 240 × 240) of a patient is input from the test dataset, to produce pixel-wise segmentation of the tumor regions and the background.

Quantitative metrics used for evaluating the segmentation results (P) w.r.t. the ground truth (G) (in case of training) and through the Leaderboard/blind testing (in case of validation)

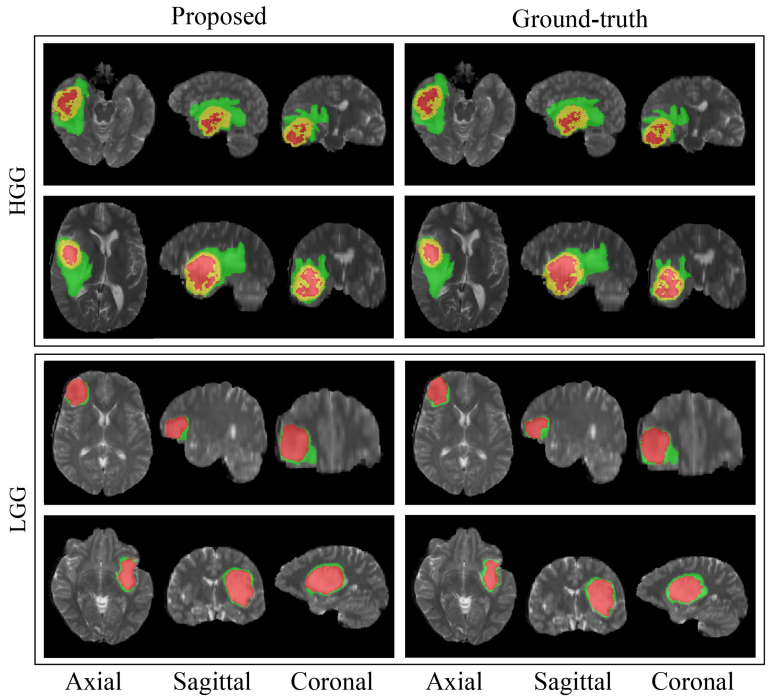


FIGURE 7 | Sample segmentation results for four patients from the BraTS 2018 training dataset. The green label is edema, the red label is non-enhancing or necrotic tumor core, and the yellow label is enhancing tumor core.

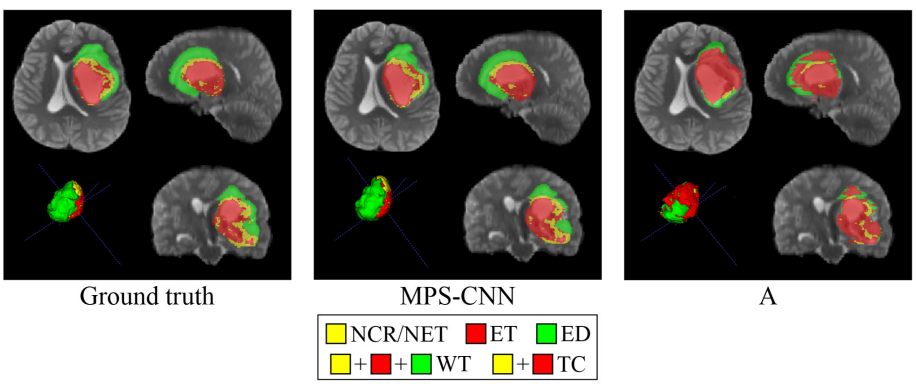


FIGURE 8 | Comparative study on segmentation obtained by our model MPS-CNN, with respect to the ground truth and Model A, for a sample patient (PID: BraTS18_2013_11_1).

are (i) Dice score $= \left(\frac{2|P_1 \bigwedge G_1|}{|P_1|+|G_1|}\right)$, (ii) sensitivity $= \left(\frac{|P_1 \bigwedge G_1|}{|G_1|}\right)$, (iii) specificity $= \left(\frac{|P_0 \bigwedge G_0|}{|G_0|}\right)$, and (iv) Hausdorff distance $= \max\{\sup_{p\in\partial P1} \inf_{g\in\partial G1} d(p,g), \sup_{g\in\partial G1} \inf_{p\in\partial P1} d(g,t)\}$, computed for WT, TC, and ET (Menze et al., 2015). Here voxels with label 0 and 1 are denoted by P_0/T_0 and P_1/T_1, respectively. The Hausdorff distance computes maximum of the shortest least-square distance d, between all points on the surfaces $\partial P1$ and $\partial G1$ of the two volumes $P1$ and $G1$.

We performed two experiments to analyze (a) the effect on performance improvement through the proposed modifications in the vanilla FCN structure, and (b) the effect of the proposed aggregated loss function in terms of handling class imbalance. The hyperparameters, employed through all the experiments, are provided in **Table 1**. These were selected through automatic cross-validation of the baseline model. Since deep CNNs entail a large number of free trainable parameters, the effective number of training samples were artificially enhanced using real time data augmentation in the form of linear transformation like random rotation (0–10°), horizontal and vertical shifts, horizontal and vertical flips. A small part of the training set (20%) was used for validating the ConvNet model, after each training epoch, for parameter selection and detection of overfitting. Each model was trained for 20 epochs, with a single epoch consuming about an

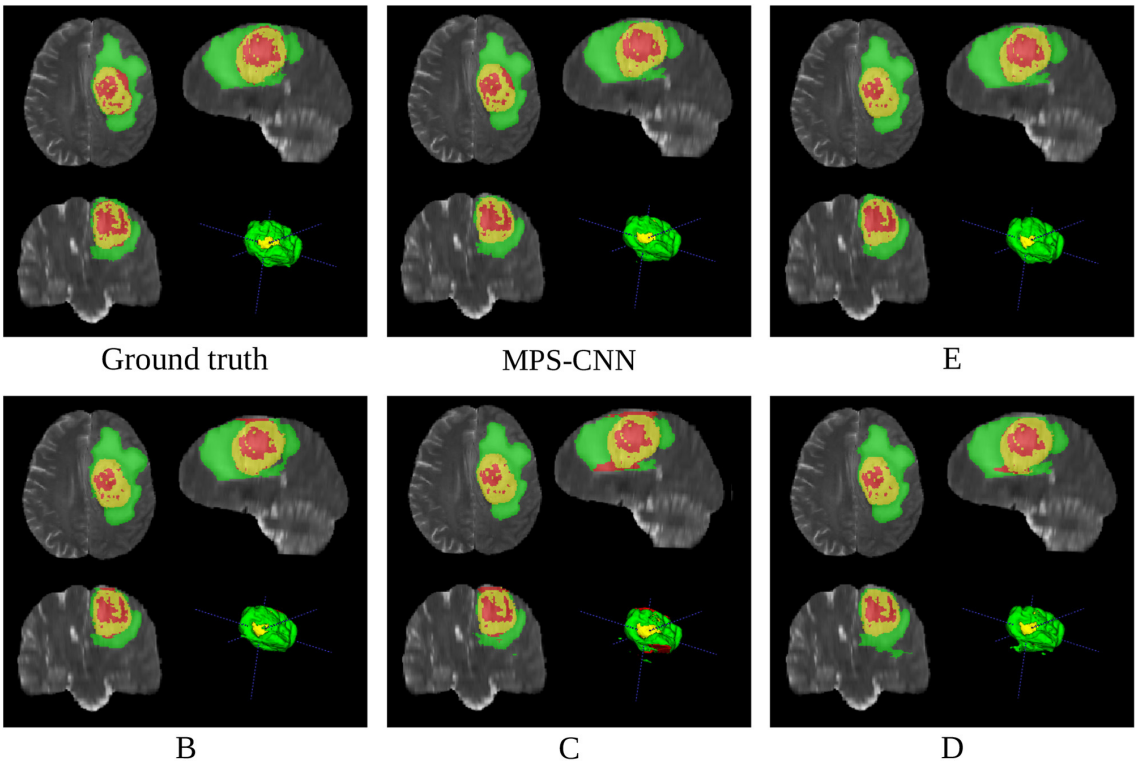


FIGURE 9 | Comparative study on segmentation obtained by our model MPS-CNN, with respect to the ground truth and Models B–E, for a sample patient (PID: PID: BraTS18_2013_7_1).

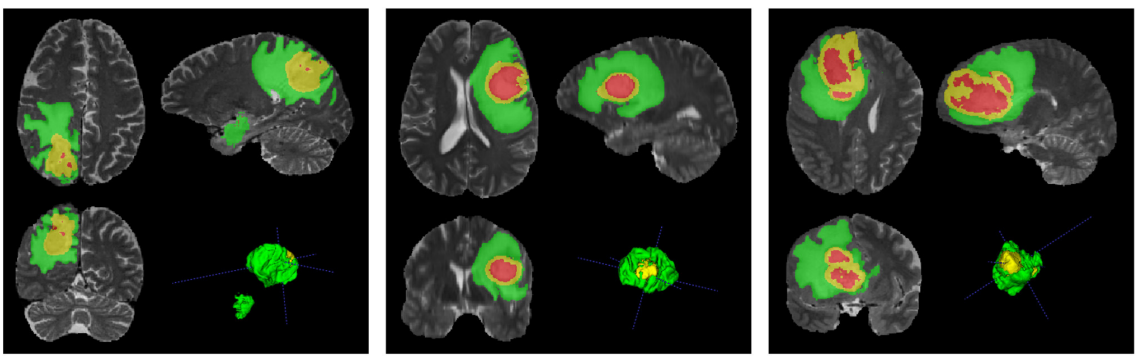


FIGURE 10 | Segmentation results obtained by Model MPS-CNN on the validation dataset for three sample patients (PIDs: BraTS18_CBICA_AAM_1, BraTS18_CBICA_ALZ_1, and BraTS18_CBICA_AUE_1).

hour (approximately) on Intel AI DevCloud platform. Inference time, including 3D CRF based refinement, required about 10 min per patient (approximately).

3.1. Experiment 1

The proposed model MPS-CNN was compared with ten variants, as outlined below.

- **Model A**: Replacing the spatial-max-pooling and max-unpooling layers of the MPS-CNN by normal max-pooling and upsampling layers.
- **Models B–D**: Architectures same as MPS-CNN, but without incorporating multi-planar aggregation and CRF based post-processing. Models B, C, and D were trained by patches, extracted (respectively) along axial, sagittal, or coronal plane only.
- **Model E**: MPS-CNN model excluding only the CRF based post-refinement.
- **Models F–J**: Training MPS-CNN with unweighted [Equation (4) with $wa_c = 1$] and weighted dice loss (Equation 4) to generate models F and G. Next unweighted [Equation (5) with

TABLE 2 | Comparative performance of MPS-CNN (radiomics-miu) with the top five models on the BraTS 2018 leader board ("NVDLMED," "SCUT_EE_CSC," "SHealth," "MIC-DKFZ," and "SUSTech").

Metric	Team	Radiomics-miu MPS-CNN	NVDLMED	SCUT_EE_CSC	SHealth	MIC-DKFZ	SUSTech
Dice	ET	0.82445	0.82531	0.81079	0.81544	0.80871	0.80522
	WT	0.90216	0.91205	0.9052	0.91204	0.91257	0.90444
	TC	0.87247	0.87049	0.85534	0.85647	0.86337	0.84943
Sensitivity	ET	0.86909	0.84497	0.83177	0.85053	0.83115	0.83064
	WT	0.91372	0.92311	0.92345	0.91968	0.91872	0.90688
	TC	0.87359	0.86405	0.87227	0.85235	0.84443	0.83156
Specificity	ET	0.99742	0.99791	0.99790	0.99753	0.99792	0.99815
	WT	0.99329	0.99519	0.99404	0.99474	0.99546	0.99549
	TC	0.99727	0.99823	0.9976	0.99773	0.99860	0.99863
Hausdorff95	ET	2.63608	3.99705	2.5551	4.04612	2.41312	2.77719
	WT	4.74851	4.5373	4.10453	4.23619	4.26797	6.32753
	TC	5.06124	6.76133	7.17313	7.21809	6.51823	6.37318

Top three scores are marked in red, green, and blue colors, respectively.

$ws_c = 1$] and weighted log loss (Equation 5) were considered to formulate models H and I. Model J was designed by training MPS-CNN with multiclass Focal loss (Lin et al., 2017), which was developed for addressing massive class imbalance.

Different models were compared based on their segmentation performance on the validation dataset, for which the organizers did not share the tumor grade (HGG/LGG) or the ground truth segmentation. During testing, the participants were required to upload the segmentation masks generated by their algorithm to the dedicated server https://www.cbica.upenn.edu/BraTS18/ for evaluation.

The box-and-whisker plots in **Figure 6** report the Dice score and Hausdroff performance of the segmentation result for the nested tumor sub-regions WT, TC, and ET for the 66 patients from BraTS 2018 validation dataset for the MPS-CNN as well as the other ten (A–J) models. The plots report the minimum & maximum; lower, median, upper quartiles; mean Dice and Hausdorff scores. The mean is marked by a red square in each case. Student's *t*-test is used to check whether the performance difference between the proposed MPS-CNN and each of the other ten compared models (A–J) is statistically significant based on their Dice score. It is evident from **Figure 6** that the proposed MPS-CNN achieved the best Dice score (Dice) and Hausdorff distance (HD) for all the three tumor sub-regions (viz. ET, TC, and WT). **Figure 7** demonstrates the segmentation obtained by our model MPS-CNN with reference to the corresponding ground truth, for two sample HGG and LGG patients from the training dataset.

Figures 8, **9** present a comparative study on the qualitative segmentation results by our model MPS-CNN and models A-E (as outlined above), to visualize the effect of the proposed modifications with respect to the basic FCN architecture. This serves to highlight the effect of the novel concepts of spatial-max-pooling and unpooling layers, along with that of multiplanar aggregation through visual demonstration on sample patients from the training dataset along all three planes (viz. axial, sagittal, coronal). Each figure also displays the ground truth segmentation. It is visually evident from **Figure 9** that segmentation by model A suffers from misclassification error along the boundary of the different tumor sub-regions, with gross error in segmenting the small sub-region *ET*. On the other hand, our model MPS-CNN produced comparable segmentation w.r.t. the ground truth, for each of the tumor sub-regions.

Figure 9 demonstrates the role of multiplanar aggregation and CRF based post-processing for a sample patient. The first row presents segmentation results obtained with multiplanar aggregation with (and without) CRF based post-processing by the models MPS-CNN (and E), respectively, with reference to the corresponding ground truth. The second row illustrates segmentation by models trained on patches extracted only along a single anatomical plane (axial, sagittal, and coronal), corresponding to models B, C, D, respectively. It is clearly observed that the aggregated models, MPS-CNN and E, perform better than any of B, C, D which were trained only along a single plane. Besides, the CRF based post-processing helps MPS-CNN to achieve more structured predictions by retaining the local and

contextual consistency. Thereby, some of the isolated NCR/NET regions get correctly segmented by our MPS-CNN as compared to Model E.

Figure 10 depicts the segmentation results, obtained by our MPS-CNN, on the validation dataset provided for three sample patients. Incidentally the models F, G, which were trained using unweighted versions of dice and log losses, were found to perform the worst due to the problem of class imbalance (as discussed in section 2.2.5). The performance gradually improved by introducing class weights to the loss functions in models H and I. However, the Focal loss function is observed to perform well in handling intra-class imbalance (for example, the amount of ET in the TC is not the same for HGG and LGG patients). However, it is less useful for cases involving inter-class imbalance.

3.2. Experiment 2

Our proposed model (MPS-CNN) was next compared with the top five models (based on the leaderboard performance on the validation dataset) that participated in the BraTS 2018 challenge, available online at (https://www.cbica.upenn.edu/BraTS18/lboardValidation.html). The name of our team is "radiomics-miu" and the other five teams selected for the comparison are "NVDLMED," "SCUT_EE_CSC," "SHealth," "MIC-DKFZ," and "SUSTech." Segmentation performance of each model is measured in terms of "Dice score," "Sensitivity," "Specificity," and "Hausdorff distance" (Menze et al., 2015). Three colors (red, blue, and green) are used to mark the first, second, and third highest scores, respectively (for each measure), as reported in **Table 2**.

It is observed that our model MPS-CNN attained the highest scores in five comparisons. It performed the best for *ET* and *TC* segmentation tasks, as compared to its nearest competitor ("NVDLMED") in terms of both the quantitative measures (Dice and Hausdorff). It is to be noted that the segmentation of *ET* and *TC* is challenging, and our MPS-CNN consistently performed best for both these tasks. In case of the *WT* segmentation it also acquired the second best accuracy, with a score which was only 1% less than that of the best performing method.

4. CONCLUSIONS

Manual segmentation of tumors from MRI is a highly tedious, time-consuming and error-prone task, mainly due to factors, such as human fatigue, overabundance of MRI slices per patient, and an increasing number of patients. Such manual operations often lead to inaccurate delineation. Development of automated and reproducible methodologies for accurate brain tumor segmentation is likely to have great clinical impact, since automated decision-making reduces human bias and is faster. We have developed a deep learning based model called Multi-Planar Spatial Convolutional Neural Network (MPS-CNN), for the automated segmentation of brain tumors from multi-modal MR images. The encoder-decoder type ConvNet model for pixel-wise segmentation was found to perform better than other patch-based models, mainly due to the introduction of new concepts like spatial max-pooling and unpooling to preserve the spatial locations of the edge pixels while reducing segmentation error around the boundaries. Integrated prediction from multiple anatomical planes (axial, sagittal, and coronal) was superior, in terms of accuracy and robustness of decision (as the data comes from multiple sources), with respect to the estimation based on any single plane. Shortcut connections were also incorporated to copy and concatenate the receptive fields, from the encoder to the decoder parts, to help the decoder network localize and recover the object details more efficiently. Very high segmentation scores were obtained on the test dataset in the blind testing phase. The effectiveness of the proposed aggregated loss function was demonstrated in terms of handling data imbalance, and the MPS-CNN model was found to be perform the best for the smaller classes viz. ET and TC. The CRF based post-refinement enhanced the segmentation accuracy by eliminating false positive regions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2018/data.html.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Multimodal Brain Tumor Segmentation Challenge 2018. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SB conceived the experiments, conducted the experiments, analyzed the results, and wrote the manuscript with support from SM. All authors discussed the results and contributed to the final manuscript.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of Intel Corporation for providing access to the Intel AI DevCloud platform used in this work.

SB acknowledges the support provided to him by the Intel Corporation, through the Intel AI Student Ambassador Program.

This publication is an outcome of the R&D work undertaken project under the Visvesvaraya Ph.D. Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation.

SM acknowledges the support provided to her by the Indian National Academy of Engineering, through the INAE Chair Professorship.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Arch.* doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017c). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. *Cancer Imaging Arch.* doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). 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.*

Banerjee, S., Mitra, S., and Uma Shankar, B. (2016a). Single seed delineation of brain tumor using multi-thresholding. *Inform. Sci.* 330, 88–103. doi: 10.1016/j.ins.2015.10.018

Banerjee, S., Mitra, S., and Uma Shankar, B. (2017). "Synergetic neuro-fuzzy feature selection and classification of brain tumors," in *Proceedings of IEEE International Conference on Fuzzy Systems* (Naples: FUZZ-IEEE), 1–6.

Banerjee, S., Mitra, S., and Uma Shankar, B. (2018a). Automated 3D segmentation of brain tumor using visual saliency. *Inform. Sci.* 424, 337–353. doi: 10.1016/j.ins.2017.10.011

Banerjee, S., Mitra, S., and Uma Shankar, B. (2018b). "Multi-planar spatial-ConvNet for segmentation and survival prediction in brain cancer," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 94–104.

Banerjee, S., Mitra, S., Uma Shankar, B., and Hayashi, Y. (2016b). A novel GBM saliency detection model using multi-channel MRI. *PLoS ONE* 11:e0146388. doi: 10.1371/journal.pone.0146388

Bauer, S., Nolte, L.-P., and Reyes, M. (2011). "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011* (Toronto, ON: Springer), 354–361.

Bauer, S., Wiest, R., Nolte, L.-P., and Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* 58, 97–129. doi: 10.1088/0031-9155/58/13/R97

Cha, S. (2006). Update on brain tumor imaging: from anatomy to physiology. *Am. J. Neuroradiol.* 27, 475–487.

Cuadra, M. B., Pollo, C., Bardera, A., Cuisenaire, O., Villemure, J.-G., and Thiran, J.-P. (2004). Atlas-based segmentation of pathological MR brain images using a model of lesion growth. *IEEE Trans. Med. Imaging* 23, 1301–1314. doi: 10.1109/TMI.2004.834618

DeAngelis, L. M. (2001). Brain tumors. *N Engl J Med.* 344, 114–123. doi: 10.1056/NEJM200101113440207

Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1915–1929. doi: 10.1109/TPAMI.2012.231

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics* (Sardinia), 249–256.

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. (2013). Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv [Preprint]. arXiv: 1312.6082.*

Gooya, A., Pohl, K. M., Bilello, M., Cirillo, L., Biros, G., Melhem, E. R., et al. (2012). GLISTR: glioma image segmentation and registration. *IEEE Trans. Med. Imaging* 31, 1941–1954. doi: 10.1109/TMI.2012.2210558

Gordillo, N., Montseny, E., and Sobrevilla, P. (2013). State of the art survey on MRI brain tumor segmentation. *Magn. Reson. Imaging* 31, 1426–1438. doi: 10.1016/j.mri.2013.05.002

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2018). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 450–462.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., et al. (2016). "Deepmedic for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, B. Menze, O. Maier, M. Reyes, S. Winzeck, and H. Handels (Cham: Springer International Publishing), 138–149.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]. arXiv: 1412.6980.*

Krähenbühl, P., and Koltun, V. (2011). "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems* (Granada), 109–117.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Pprocessing Systems* (Nevada, CA), 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Li, Y., Wang, D., Wang, L., Yu, J., Du, D., Chen, Y., et al. (2013). Distinct genomics aberrations between low-grade and high-grade gliomas of Chinese patients. *PLoS ONE* 8:e57168. doi: 10.1371/journal.pone.0057168

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 2980–2988.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BraTS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Menze, B. H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., and Golland, P. (2010). "A generative model for brain tumor segmentation in multi-modal images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010* (Beijing: Springer), 151–159.

Mitra, S., Banerjee, S., and Hayashi, Y. (2017). Volumetric brain tumour detection from MRI using visual saliency. *PLoS ONE* 12:e0187209. doi: 10.1371/journal.pone.0187209

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* 35, 1240–1251. doi: 10.1109/TMI.2016.2538465

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: integrated recognition, localization and detection using convolutional networks. *arXiv [Preprint]. arXiv: 1312.6229.*

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

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Québec City, QC: Springer), 240–248.

Urban, G., Bendszus, M., Hamprecht, F. A., and Kleesiek, J. (2014). "Multi-modal brain tumor segmentation using deep convolutional neural networks," in *Proceedings of MICCAI-BRATS* (Boston, MA: Winning Contribution), 1–5.

Wu, W., Chen, A. Y. C., Zhao, L., and Corso, J. J. (2014). Brain tumor detection and segmentation in a CRF (conditional random fields) framework with pixel-pairwise affinity and superpixel-level features. *Int. J. Comput. Assist. Radiol. Surg.* 9, 241–253. doi: 10.1007/s11548-013-0922-7

Zacharaki, E. I., Shen, D., Lee, S.-K., and Davatzikos, C. (2008). ORBIT: a multiresolution framework for deformable registration of brain tumor images. *IEEE Trans. Med. Imaging* 27, 1003–1017. doi: 10.1109/TMI.2008.916954

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., et al. (2012a). "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012* (Nice: Springer), 369–376.

Zikic, D., Glocker, B., Konukoglu, E., Shotton, J., Criminisi, A., Ye, D., et al. (2012b). "Context-sensitive classification forests for segmentation of brain tumor tissues," in *Proceedings of MICCAI-BraTS* (Nice), 22–30.

Zikic, D., Ioannou, Y., Brown, M., and Criminisi, A. (2014). "Segmentation of brain tumor tissues with convolutional neural networks," in *Proceedings of MICCAI-BRATS* (Boston, MA), 36–39.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 24 January 2020
doi: 10.3389/fnins.2019.01449

Improving Patch-Based Convolutional Neural Networks for MRI Brain Tumor Segmentation by Leveraging Location Information

Po-Yu Kao [1], Shailja Shailja [1], Jiaxiang Jiang [1], Angela Zhang [1], Amil Khan [1], Jefferson W. Chen [2] and B. S. Manjunath [1]**

[1] Vision Research Lab, Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA, United States, [2] Department of Neurological Surgery, University of California, Irvine, Irvine, CA, United States

The manual brain tumor annotation process is time consuming and resource consuming, therefore, an automated and accurate brain tumor segmentation tool is greatly in demand. In this paper, we introduce a novel method to integrate location information with the state-of-the-art patch-based neural networks for brain tumor segmentation. This is motivated by the observation that lesions are not uniformly distributed across different brain parcellation regions and that a locality-sensitive segmentation is likely to obtain better segmentation accuracy. Toward this, we use an existing brain parcellation atlas in the Montreal Neurological Institute (MNI) space and map this atlas to the individual subject data. This mapped atlas in the subject data space is integrated with structural Magnetic Resonance (MR) imaging data, and patch-based neural networks, including 3D U-Net and DeepMedic, are trained to classify the different brain lesions. Multiple state-of-the-art neural networks are trained and integrated with XGBoost fusion in the proposed two-level ensemble method. The first level reduces the uncertainty of the same type of models with different seed initializations, and the second level leverages the advantages of different types of neural network models. The proposed location information fusion method improves the segmentation performance of state-of-the-art networks including 3D U-Net and DeepMedic. Our proposed ensemble also achieves better segmentation performance compared to the state-of-the-art networks in BraTS 2017 and rivals state-of-the-art networks in BraTS 2018. Detailed results are provided on the public multimodal brain tumor segmentation (BraTS) benchmarks.

Keywords: gliomas, brain tumor segmentation, brain parcellation atlas, convolutional neural network, DeepMedic, 3D U-Net, ensemble learning, XGBoost

OPEN ACCESS

Edited by:
Spyridon Bakas, University of Pennsylvania, United States

Reviewed by:
Yi Su, Banner Alzheimer's Institute, United States
Anahita Fathi Kazerooni, University of Pennsylvania, United States

****Correspondence:***
Po-Yu Kao
poyu_kao@ucsb.edu
B. S. Manjunath
manj@ucsb.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *26 August 2019*
Accepted: *27 December 2019*
Published: *24 January 2020*

Citation:
Kao P-Y, Shailja S, Jiang J, Zhang A, Khan A, Chen JW and Manjunath BS (2020) Improving Patch-Based Convolutional Neural Networks for MRI Brain Tumor Segmentation by Leveraging Location Information. Front. Neurosci. 13:1449. doi: 10.3389/fnins.2019.01449

1. INTRODUCTION

Glioma is a common type of brain tumor in adults originating in the glial cells that support neurons and help them function. The World Health Organization (WHO) classification system categorizes gliomas from grade I (lowest grade) through grade IV (highest grade), based upon histopathologic characteristics that predict their behavior over time (Louis et al., 2007). Low-grade gliomas (LGGs) consist of WHO-grade I tumors and WHO-grade II tumors, that tend to exhibit benign tendencies

and indicate a better prognosis for the patient. WHO-grade III and IV tumors are included in high-grade gliomas (HGG) that are malignant and more aggressive. Patients with HGG had median survival time (MST) 18 months, and the MST of patients with Grade III and IV glioma were 26 and 13 months, respectively (Noiphithak and Veerasarn, 2017). Gliomas are further divided into four types of sub-regions, namely edema, non-enhancing core, necrotic core, and enhancing core based on the acuteness of the tumor cells that have different appearances in MR imaging data. However, segmenting the different sub-regions of gliomas is a daunting task because of the intrinsic heterogeneity which affects their visual appearance as well as shape. Clinically, MR images help a doctor to evaluate the tumor and plan treatment. Moreover, the treatment depends on the type, size, shape, grade, and location of the tumor, which varies widely. Consequently, this observation leads to the importance of an accurate brain tumor segmentation for better diagnosis of brain tumors. Also, the manual annotation process is time consuming and resource consuming, therefore, an automated and accurate brain tumor segmentation tool is greatly in demand.

Deep neural networks (DNNs) have achieved state-of-the-art segmentation performance on the recent Multimodal Brain Tumor Segmentation (BraTS) Challenges (Bakas et al., 2018). Kamnitsas et al. (2017a) conducted the comparative study on performance and concluded that deep learning along with ensemble learning-based methods outperform the others as they leverage the advantage of each deep learning model. Wang et al. (2017) analyzed three different binary segmentations task rather than a single multi-class segmentation task, and three different binary segmentations task has a better performance than a single multi-class segmentation task. Along this line, Isensee et al. (2017) proposed to integrate segmentation layers at different levels of optimized 3D U-Net-like architectures followed by element-wise summation. Myronenko (2018) implemented a modified decoder and encoder structure of CNN to generate dense segmentation. Likewise, Isensee et al. (2018) demonstrated that an original U-Net architecture trained with additional institution dataset improved the dice score of enhancing tumor. McKinley et al. (2018) also proposed a U-Net-like network and introduce a new loss function, a generalization of binary cross-entropy, to account for label uncertainty. Furthermore, Zhou et al. (2018) explored the ensemble of different networks including multi-scale context information, and also segmented three tumor compartments in cascade with an additional attention block.

Our recent work (Kao et al., 2018) utilizes an existing parcellation to bring location information of the brain into patch-based neural networks that improve the brain tumor segmentation performance of networks. Outputs from 26 models were averaged, including 19 different types of DeepMedics (Kamnitsas et al., 2017b) and seven different types of 3D U-Nets (Çiçek et al., 2016), to get the final tumor predictions. Different from our previous ensemble, the proposed ensemble only contains six models including three DeepMedics and three 3D U-Nets with different seed initializations that only take <1 min in the inference time. We also propose a novel two-level ensemble method which reduces the uncertainty of predictions in the first level and takes advantage of different types of models in the second level. In this paper, we also demonstrate that the proposed location fusion methods improve the segmentation performance of the single state-of-the-art patch-based network and an ensemble of multiple state-of-the-art patch-based networks. The proposed ensemble has better segmentation performance compared to state-of-the-art networks in BraTS 2017 dataset and competitive performance to the state-of-the-art networks in BraTS 2018 dataset. The main contribution of this paper is two-fold. First, it proposes a location information fusion method that improves the segmentation performance of state-of-the-art networks including DeepMedic and 3D U-Net. Second, it proposes a novel two-level ensemble method which reduces the uncertainty of prediction and leverages the advantages of different segmentation networks.

2. MATERIALS AND METHODS

This section describes the details of (i) a proposed location information fusion method for improving brain tumor segmentation using a patch-based convolutional neural network (CNN), and (ii) a proposed ensemble learning method which takes advantage of model diversity and uncertainty reduction. This section includes the data description, data pre-processing, network architectures, training, and test procedure, proposed location information fusion method, and proposed ensemble methods. The evaluation metrics are also described at the end of this section.

2.1. Dataset

The Multimodal Brain Tumor Segmentation Challenges (BraTS) 2017 dataset and BraTS 2018 dataset (Menze et al., 2015; Bakas et al., 2017a,b,c) comprise clinically-acquired pre-operative multimodal MRI scans of glioblastoma (GBM/HGG) and lower-grade glioma (LGG) as training, validation and test data. There are 285 subjects in the training set and 46 and 66 subjects in the validation set of BraTS 2017 and BraTS 2018, respectively. The lesion ground-truth labels are available for the training subjects but withheld for both the validation and test subjects. MRI scans were available as native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR) volumes. These scans were distributed after being skull-stripped, pre-processed, re-sampled, and interpolated into 1 mm isotropic resolution with an image size of $240 \times 240 \times 155$ in x-, y-, and z-direction. Tumor segmentation labels were produced manually by a trained team of radiologists and radiographers. The edema was segmented primarily from T2 images, non-enhancing and enhancing the core of the tumor from T1c together with the lesions visible in T1 and necrotic core from T1c. We used the annotated and co-registered imaging datasets including the Gd-enhancing tumor, the peri-tumoral edema and the necrotic and non-enhancing tumor core for our training and test procedure.

2.2. Data Pre-processing

Different modalities used for mapping tumor-induced tissue changes include MR-T1, MR-T1Gd, MR-T2, and MR-FLAIR,

which leads to varying intensity ranges. We first normalize each modality to a standard range of values. Each MR image is pre-processed by first clipping it at (0.2 percentile, 99.8 percentile) of non-zero voxels to remove the outliers. Subsequently, each modality is normalized individually using $\bar{x}_i = (x_i - \mu)/\sigma$ where i is the index of voxel inside the brain, $\bar{x}_i$ is the normalized voxel, x_i is the corresponding raw voxel, and μ and σ are the mean and standard deviation of the raw voxels inside the brain, respectively.

2.3. Network Architectures

Two different network architectures adapted from DeepMedic (Kamnitsas et al., 2017b) and 3D U-Net (Çiçek et al., 2016) are examined in this study. DeepMedic was initially designed for brain lesion segmentation, e.g., stroke lesions (Kamnitsas et al., 2015) and brain tumor lesions (Kamnitsas et al., 2016), and 3D U-Net which is the 3D version of U-Net (Ronneberger et al., 2015) is widely used for the volumetric image segmentation tasks (Yu et al., 2017; Li et al., 2018; Jiang et al., 2019). More details of network architectures are described below.

2.3.1. Modified DeepMedic

The first network architecture shown in **Figure 1** is modified from DeepMedic (Kamnitsas et al., 2017b). The number of convolutional kernels is indicated within the white box. Batch normalization (Ioffe and Szegedy, 2015) is used. Residual connection (He et al., 2016) is used in the normal resolution path, and trilinear interpolation is used in the upsampling layer of the downsampled resolution path. The size of the receptive field of the normal resolution path is 25 × 25 × 25, and the size of the receptive field of the downsampled resolution path is 19 × 19 × 19. The receptive field of downsampled resolution path is downsampled from an image patch of size 55 × 55 × 55 by a factor of 3 in the same center as the receptive field of normal resolution path. The modified DeepMedic predicts the central 9 × 9 × 9 voxels of the receptive field of normal resolution path.

2.3.1.1. Training and test procedure

The modified DeepMedic is only trained with patches that have approximately 50% foreground (lesion) and 50% background to solve the class imbalance problem, and it is trained with batch size 50. In every epoch, 20 patches are extracted from each subject. The network is trained for a total of 500 epochs. The weights of the network are updated by Adam algorithm (Kingma and Ba, 2015) with an initial learning rate of $l_0 = 10^{-3}$ following the schedule of $l_0 \times 0.1^{\text{epoch}}$, L2 penalty weight decay of 10^{-4}, and AMSGrad (Reddi et al., 2018). A standard multi-class cross-entropy loss is used. Randomly flipping in x-, y-, and z-axis with a probability of 50%, and random noise are applied in the data augmentation of the training procedure. At the test time, a sliding window scheme of step size 9 is used to get the tumor lesion prediction

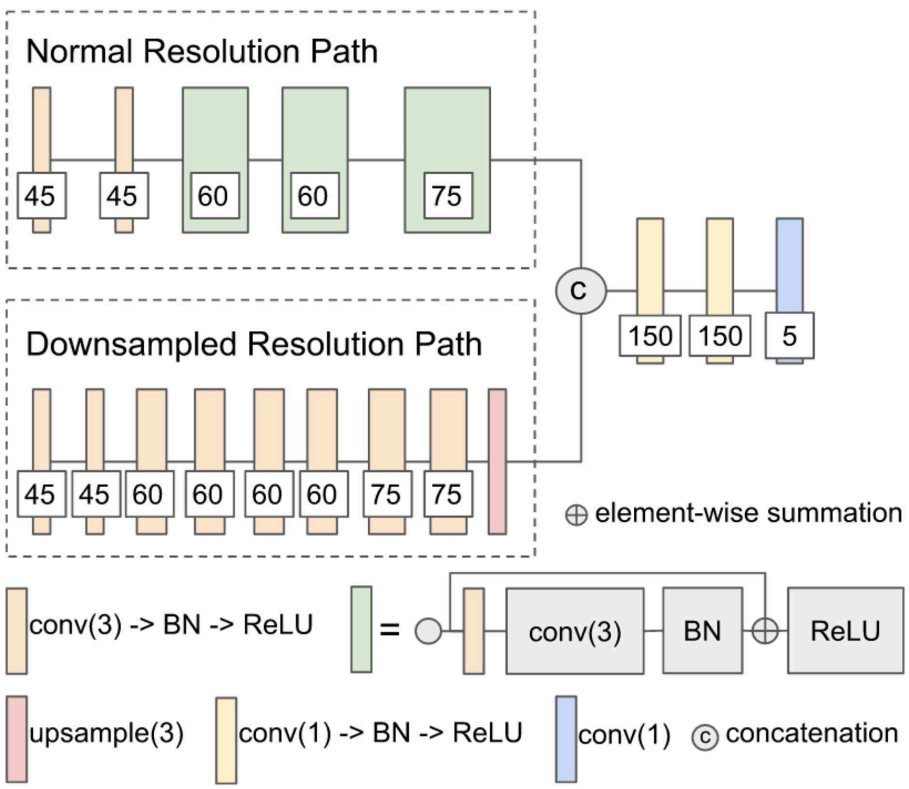


FIGURE 1 | The network architecture of modified DeepMedic. conv(3), 3 × 3 × 3 convolutional layer; BN, batch normalization; upsample(3), trilinear interpolation by a factor of 3; and conv(1), 1 × 1 × 1 convolutional layer.

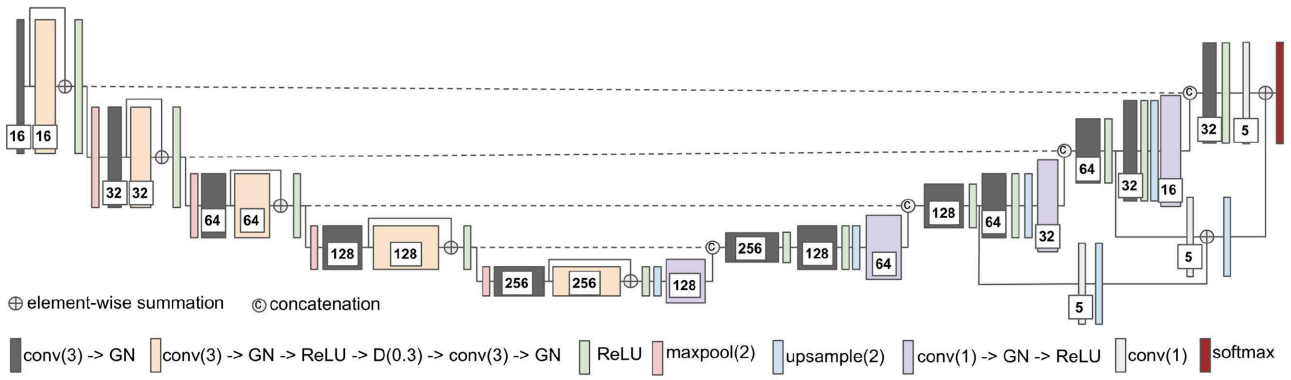


FIGURE 2 | The network architecture of modified 3D U-Net. conv(3), 3 × 3 × 3 convolutional layer; GN, group normalization; D(0.3), dropout layer with 0.3 dropout rate; maxpool(2), 2 × 2 × 2 max pooling layer; and conv(1), 1 × 1 × 1 convolutional layer.

of the test subject. Training takes approximately 6 h, and a test for each subject takes approximately 24 s on an Nvidia 1080 Ti GPU and an Intel Xeon CPU E5-2696 v4 @ 2.20 GHz.

2.3.2. Modified 3D U-Net

The second network architecture shown in **Figure 2** is modified from 3D U-Nets (Çiçek et al., 2016). Different colors of blocks represent different types of layers. The number of convolutional kernels is indicated within the white box. Group normalization (Wu and He, 2018) is used, and the number of groups is set to 4. Residual connection (He et al., 2016) is used in the encoding path, and trilinear interpolation is used in the upsampling layer.

2.3.2.1. Training and test procedure

The modified 3D U-Net is trained with randomly cropped patches of size 128 × 128 × 128 voxels and batch size 2. In every epoch, a cropped patch is randomly extracted from each subject. The network is trained for a total of 300 epochs. The weights of the network are updated by Adam algorithm (Kingma and Ba, 2015) with an initial learning rate $l_0 = 10^{-3}$ following the schedule of $l_0 \times 0.1^{\text{epoch}}$, L2 penalty weight decay of 10^{-4}, and AMSGrad (Reddi et al., 2018). For the loss function, the standard multi-class cross-entropy loss with the hard negative mining is used to solve the class imbalance problem of the dataset. We only back-propagate the negative (background) voxels with the largest losses (hard negative) and the positive (lesions) voxels to the gradients. In our implementation, the number of selected negative voxels is at most three times more than the number of positive voxels. Besides, data augmentation is not used for both training and testing. At the test time, we input the entire image of size 240 × 240 × 155 voxels into the trained 3D U-Net for each patient to get the predicted lesion mask. Training takes approximately 12.5 h, and the test takes approximately 1.5 s per subject on an Nvidia 1080 Ti GPU and an Intel Xeon CPU E5-2696 v4 @ 2.20 GHz.

2.4. Incorporating Location Information With Patch-Based Convolutional Neural Network

The heatmaps (see **Figure 3**) of different brain tumor lesion sub-regions reveal that different lesion sub-regions have different probability occurring in different locations. The heatmaps are generated by first registering the ground-truth lesions of 285 training subjects from the subject space to the MNI 152 1mm space using FMRIB's Linear Image Registration Tool (FLIRT) (Jenkinson and Smith, 2001) from FSL, extracting the binary masks of different types of lesion sub-regions from each subject, and applying element-wise summation to the same type of binary masks of each subject in the MNI 152 1mm space. However, the patch-based convolutional neural networks (CNNs), e.g., DeepMedic or 3D U-Net, do not consider location information for brain tumor segmentation. That is, the patch-based CNNs do not know location information of the input patches.

In this study, an existing brain parcellation atlas, Harvard-Oxford Subcortical atlas (see **Figure 3**), is used as location information of the brain for the patch-based CNN. The details of Harvard-Oxford Subcortical parcellation regions are described in **Table 1**. There are two main reasons for choosing this atlas: (1) this atlas covers more than 90% of a brain region, and (2) lesion information and location information are converted into this atlas (see **Figure 3**). The distribution in **Figure 3E** is calculated by dividing the total volume of the lesion sub-regions from 285 training subjects by the total volume of the corresponding brain parcellation in the MNI 152 space. **Figure 3** shows that different lesion sub-regions have different probabilities happening in different parcellation regions.

Our proposed location information fusion method which is shown in **Figure 4** explicitly includes location information as input into a patch-based CNN. First, the Harvard-Oxford subcortical atlas is registered to the individual subject space from MNI 152 1 mm space (Grabner et al., 2006) using FLIRT (Jenkinson and Smith, 2001) from FSL. The registered atlas is then split into 21 binary masks and concatenated with the multimodal MR images as input to a patch-based CNN for both

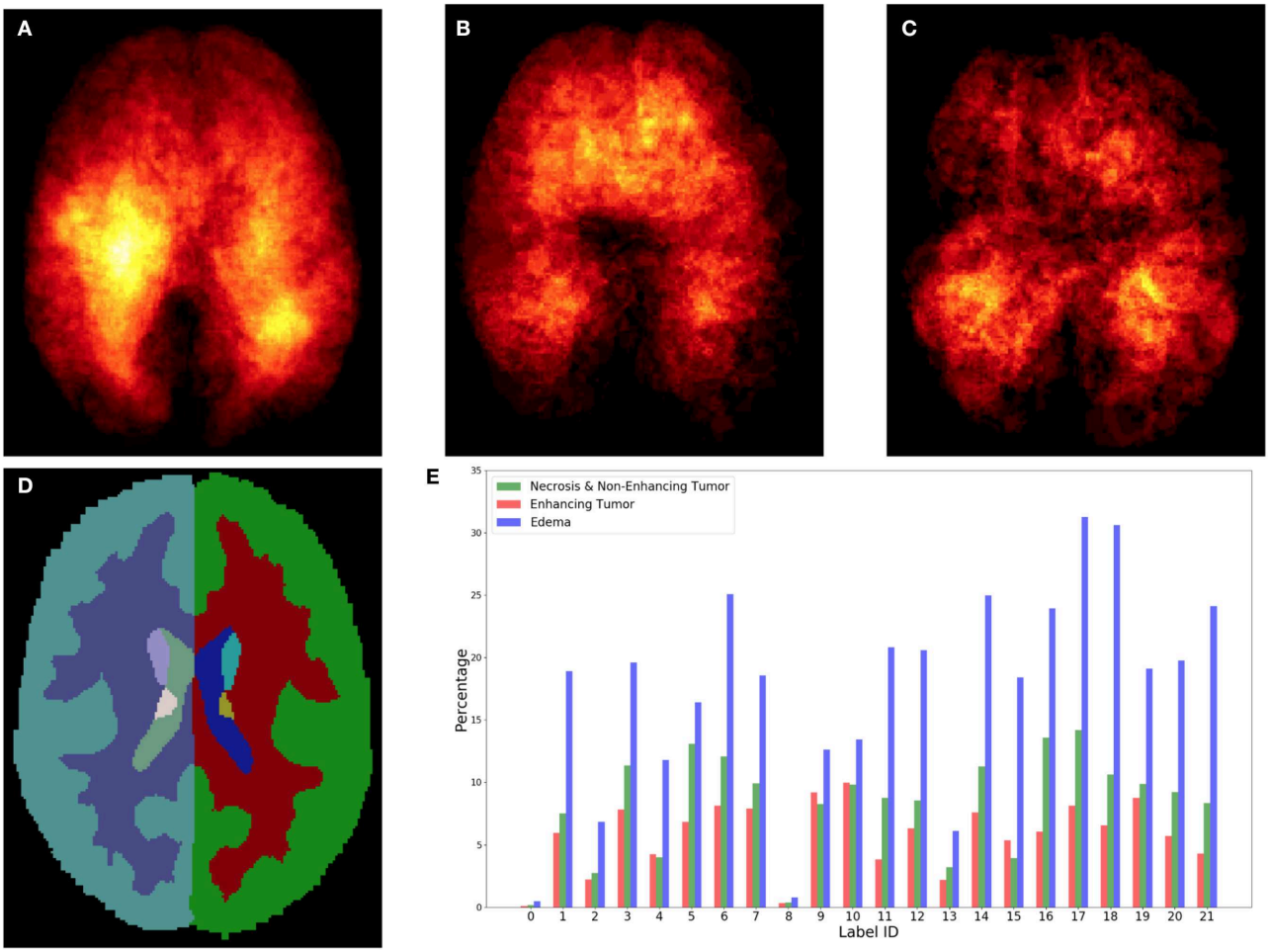


FIGURE 3 | Top row shows the heatmaps of different lesion sub-regions, **(A)**: edema, **(B)**: necrosis & non-enhancing tumor, and **(C)**: enhancing tumor, from 285 training subjects of BraTS 2018 in the MNI 152 1 mm space. The brighter (yellow) voxel represents higher value. **(D)** Shows Harvard-Oxford subcortical structural atlas (Desikan et al., 2006), and **(E)** the percentage of brain lesion sub-regions observed in different parcellation regions of the Harvard-Oxford subcortical atlas from 285 training subjects of BraTS 2018. The *x*-axis indicates the brain parcellation label ID. Regions not covered by the Harvard-Oxford subcortical atlas are in label 0.

training and test. As a result, the fused input has 25 channels. The first four channels provide the image information, and the last 21 channels contain the location information of the brain.

It is noted that the registration involving in our research only contain a linear (affline) transformation which has 9 degrees of freedom. In general, the registration should include a linear transformation followed by a deformable transformation. However, for the patient having brain lesions, a lesion mask has to be given in the deformable transformation in order to account for the effect of the lesion (Kuijf et al., 2013). The problem we have here is finding the brain tumor lesion based on the multimodal MR scan. Therefore, we are not able to use any ground truth lesion information, and the registration only contains a linear (affine) transformation.

2.5. Ensemble Methods

Ensemble methods aim at improving the predictive performance of a given statistical learning or model fitting technique. The general principle of ensemble methods is to construct a linear combination of some model fitting methods, instead of using a single fit of the method (Bühlmann, 2012). Ensembles have been proven to have better performance than any single model (Dietterich, 2000). Two-level ensemble approach, including the arithmetic mean and boosting, is proposed in this study, and more details of these methods are explained below.

2.5.1. Arithmetic Mean

The arithmetic mean, $\bar{x}$, is the average of n values $x_1, x_2, \ldots x_n$, i.e., $\bar{x} = (x_1 + x_2 + \ldots + x_n)/n$. If we have n models in our ensemble, then the arithmetic mean P is defined by the formula:

$$P = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{p_1 + p_2 + \ldots + p_n}{n} \quad (1)$$

where p_i is the probability map of model i. The arithmetic mean ensemble method reduces the uncertainties of different models.

2.5.2. XGBoost

Boosting algorithms are widely used in machine learning to achieve state-of-art performance. It improves the prediction of the models by training the base learners sequentially to improve their predecessor. There are different boosting algorithms such as AdaBoost (Freund and Schapire, 1997; Hastie et al., 2009), short for Adaptive Boosting, and Gradient Boosting (Friedman, 2001, 2002). AdaBoost tunes the weights for every incorrect classified observation at every iteration while Gradient Boosting tries to fit the new predictor to the residual errors made by the previous predictor. Both of the boosting algorithms are generally very slow in implementation and not very scalable. Chen and Guestrin (2016) described a scalable tree boosting system called XGBoost which is an implementation of gradient boosted decision trees that are efficient in run-time and space complexity. It also supports parallelization of tree construction, distributed computing for training very large models, out-of-core computing for very large datasets that do not fit into memory and cache optimization to make the best use of hardware. These features make XGBoost ideal for our purpose of study in brain tumor segmentation, therefore, it is used in our study.

TABLE 1 | The label ID and corresponding brain region of Harvard-Oxford Subcortical Atlas.

Label ID	Brain region
1	Left Cerebral White Matter
2	Left Cerebral Cortex
3	Left Lateral Ventrical
4	Left Thalamus
5	Left Caudate
6	Left Putamen
7	Left Pallidum
8	Brain-Stem
9	Left Hippocampus
10	Left Amygdala
11	Left Accumbens
12	Right Cerebral White Matter
13	Right Cerebral Cortex
14	Right Lateral Ventricle
15	Right Thalamus
16	Right Caudate
17	Right Putamen
18	Right Pallidum
19	Right Hippocampus
20	Right Amygdala
21	Right Accumbens

2.5.3. Two-Level Ensemble Approach: Arithmetic Mean and XGBoost

The ensemble of multiple identical network architectures with different seed initializations has been proven to reduce the uncertainty of models and improve the segmentation performance (Lakshminarayanan et al., 2017). Moreover, Dietterich (2000) demonstrated that the boosting algorithm has the best performance compared to bagging and randomized trees. Inspired by their works, we propose a two-level ensemble approach shown in **Figure 5** that averages the probability maps from the same type of models in the first level and then boosts the averaged probability maps from different models by using the XGBoost algorithm in the second level. We have examined three different classification strategies in the second level, and these classification strategies are based on multi-class classification and binary class classification. More details are described in sections 2.5.3.1 and 2.5.3.2.

2.5.3.1. Multi-class classification

The multi-class classification problem refers to classifying voxels into one of the four classes. It produces segmentation labels of the background and different glioma sub-regions that include: (1) the enhancing tumor, (2) the edema, and (3) the necrosis

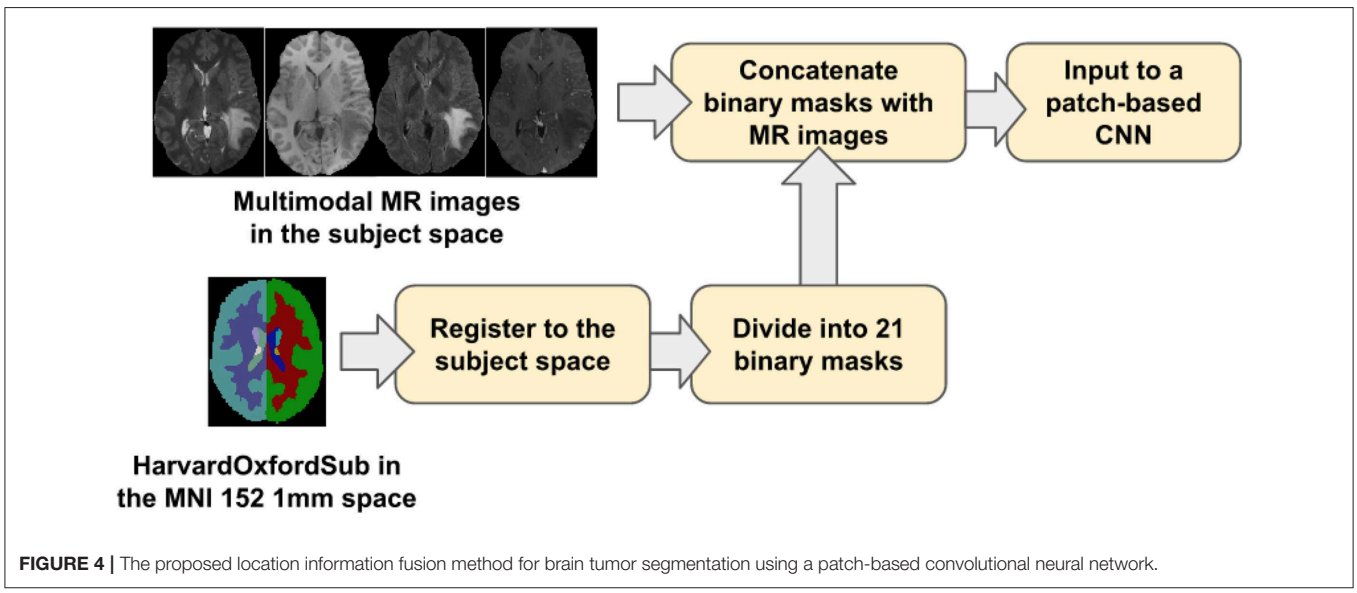


FIGURE 4 | The proposed location information fusion method for brain tumor segmentation using a patch-based convolutional neural network.

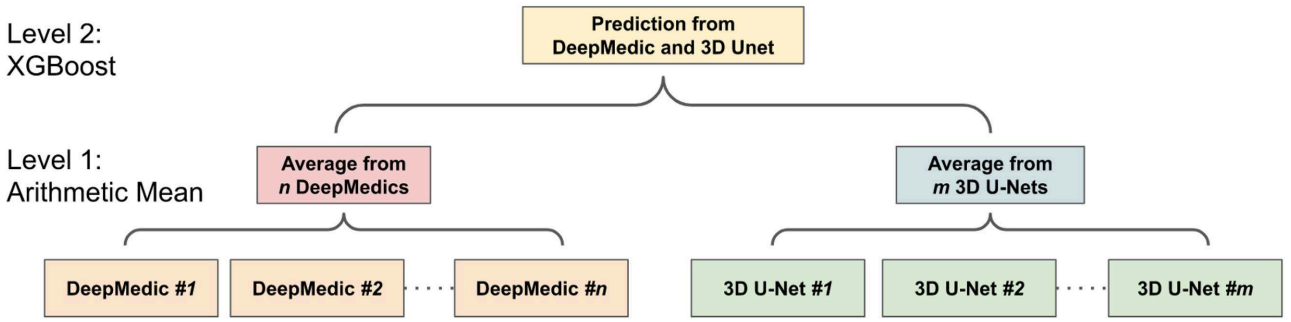


FIGURE 5 | The workflow of two-level ensemble approach.

& non-enhancing tumor. Since XGBoost is known to produce better results in different machine learning problems (Nielsen, 2016), XGBoost is used in our multi-class classification problem with the softmax function objective. The softmax function σ is defined by

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for} \quad i = 1, \ldots, K \quad \text{and}$$
$$\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K.$$

where K is the number of classes in the classification problem. Using the softmax objective function, we get a neural network that models the probability of a class z_i as multinominal distribution.

2.5.3.2. Binary classifications

The multi-class classification problem can be reduced to several binary classification problems where each binary classifier is trained to classify voxels into two classes. There are two different approaches, one-versus-all and one-versus-one, to perform such a transformation. For a k-class problem, the one-versus-all method trains k different binary classifiers where the two-class classifier C_i learns to distinguish the class i from all the other $k - i$ classes.

$$C_+ = C_i \quad \text{and} \quad C_- = \{C_j | j = 1, \cdots, K,\ j \neq i\}$$

One-vs.-one approach is based on training $k \times (k-1)/2$ classifiers, where each classifier learns to distinguish 2 classes only.

$$C_+ = C_i \quad \text{and} \quad C_- = \{C_j | j \neq i\}$$

where C_+ and C_- are the two classes of the binary class classification problem.

2.6. Evaluation Metrics

Two evaluation metrics, dice similarity score (DSC) and Hausdorff distance, are commonly used in the brain tumor segmentation problem. DSC is used to measure the similarity of the predicted lesions and ground-truth lesions, and Hausdorff distance is used to measure how far the predicted lesions are from the ground-truth lesions. More details of these two evaluation metrics are explained in the following sections.

2.6.1. Dice Similarity Score

Dice similarity score (DSC) is a statistic used to measure the similarity of two sets. It is defined as

$$DSC = \frac{2|G \cap P|}{|G| + |P|} \tag{2}$$

where $|G|$ and $|P|$ are the number of voxels in the ground-truth and prediction, respectively. DSC ranges between 0 and 1 (1 means perfect matching).

2.6.2. Hausdorff Distance

Hausdorff distance $d_H(X, Y)$ measures how far two subsets $\{X, Y\}$ of a metric space are from each other. It is defined as

$$d_H(X, Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\} \tag{3}$$

where d is the Euclidean distance, sup is the supremum, and inf is the infimum. Hausdorff distance ranges from 0 to infinity (0 means perfect matching). In this study, 95 percentile of Hausdorff distance (HD95) is used to disregard the outliers.

3. EXPERIMENTS AND RESULTS

In this section, we demonstrate the advantage of the proposed location information fusion method and the proposed two-level ensemble learning method. In Experiment 1, we first examine the segmentation performance of the proposed location information fusion method on a single model. In Experiment 2, we examine the performance of the proposed location information fusion method on an ensemble of the same type of models. In Experiment 3, we examine different ensemble methods that predict the final brain tumor lesions based on the output probability maps from DeepMedics and 3D U-Nets. In Experiment 4, we compare the segmentation performance of the proposed method with state-of-the-art methods. The details of each experiment and experimental results are described in the following sections.

TABLE 2 | Results of the first experiment on the BraTS 2018 validation set.

Model description	DSC_ET	DSC_WT	DSC_TC	HD95_ET	HD95_WT	HD95_TC
DeepMedic	78.1(25.4)	89.5(6.8)	81.4(21.3)	4.21(8.19)	10.60(15.30)	9.90(20.13)
DeepMedic + BP	**79.0(22.6)**	**89.6(6.4)**	81.3(21.8)	**3.78(7.23)**	**8.87(15.23)**	**6.55(6.81)**
3D U-Net	74.9(25.8)	89.7(7.7)	76.6(20.3)	5.85(9.50)	4.88(4.41)	10.46(13.51)
3D U-Net + BP	**76.4(25.4)**	**90.1(6.4)**	**76.9(24.4)**	**5.48(9.50)**	**4.87(6.28)**	**10.07(13.99)**

The results are reported as mean (standard deviation). Bold numbers highlight the improved results with additional brain parcellation masks within the same type of model.

TABLE 3 | Results of the second experiment on the BraTS 2018 validation set.

Ensemble description	DSC_ET	DSC_WT	DSC_TC	HD95_ET	HD95_WT	HD95_TC
DeepMedic	79.7(23.6)	90.0(6.8)	81.4(22.1)	3.94(7.77)	7.44(13.36)	8.88(14.03)
DeepMedic + BP	78.4(25.3)	**90.2(6.4)**	**81.8(21.9)**	**3.37(5.18)**	**5.64(7.53)**	**7.01(12.29)**
3D U-Net	77.6(24.2)	90.0(9.0)	78.0(21.2)	5.01(9.22)	4.39(4.05)	9.77(13.60)
3D U-Net + BP	77.4(25.1)	**90.4(6.6)**	**79.3(22.4)**	**4.25(8.31)**	4.59(6.29)	**9.66(14.20)**

The results are reported as mean (standard deviation). Bold numbers highlight the improved results with additional brain parcellation masks within the same type of ensemble.

TABLE 4 | Results of the third experiment on BraTS 2018 validation set.

Ensemble methods	DSC_ET	DSC_WT	DSC_TC	HD95_ET	HD95_WT	HD95_TC
Arith. mean	**78.3(25.4)**	90.6(6.4)	81.3(21.8)	3.72(7.90)	**4.35(6.21)**	7.77(13.45)
TLMC	78.3(25.5)	90.7(6.3)	81.0(22.1)	**2.81(3.55)**	4.38(6.26)	7.80(13.49)
TLBC	76.6(26.8)	90.7(6.2)	82.2(21.2)	7.93(2.66)	4.39(6.27)	8.34(16.98)
TLFC	78.2(25.6)	**90.8(6.1)**	**82.3(21.2)**	2.96(3.80)	4.39(6.22)	**6.91(12.64)**

The results are reported as mean (standard deviation). Bold numbers highlight the best performance between different ensemble methods.

3.1. Experiment 1: Location Information Fusion Method on a Single Model

In the first experiment, we would like to examine the performance of the proposed location information fusion method on a single patch-based neural network. We first train a DeepMedic and a 3D U-Net using only multimodal MR images. Thereafter, we train another identical DeepMedic and another identical 3D U-Net with multimodal MR images and binary brain parcellation masks. BraTS 2018 training set is used to train the models with five-fold cross-validation, and the BraTS 2018 validation set is used as the test set. The experimental results are shown in **Table 2**.

3.2. Experiment 2: Location Information Fusion Method on an Ensemble

In the second experiment, we would like to examine the performance of the proposed location information fusion on the ensemble of DeepMedics and the ensemble of 3D U-Nets. Each ensemble has identical network architectures with different seed initializations, and the output of the ensemble is the arithmetic mean from networks. We first train ensembles of DeepMedics without additional brain parcellation masks. Thereafter, we train ensembles of 3D U-Nets without additional brain parcellation masks. In the end, we train another identical ensemble of DeepMedics and another identical ensemble of 3D U-Nets with additional brain parcellation masks. BraTS 2018 training set is used to train the models with five-fold cross-validation, and the BraTS 2018 validation set is used as the test set. The experimental results are shown in **Table 3**.

3.3. Experiment 3: Different Ensemble Methods

In the third experiment, we would like to exam the performance of different ensemble methods including arithmetic mean and two-level ensemble approaches described in section 2.5. We first train three identical DeepMedics with additional brain parcellation channels and different seed initializations. We also train three identical 3D U-Nets with additional brain parcellation channels and different seed initializations. Then, we apply different ensemble methods on the probability maps from these models to generate the final tumor segmentation mask. More details of different ensemble methods are described below.

3.3.1. Experiment 3.1: Arithmetic Mean

In this experiment, the final tumor segmentation mask is directly generated by averaging the probability maps from three DeepMedics and three 3D U-Nets. BraTS 2018 training set is used to train the models with five-fold cross-validation, and the BraTS 2018 validation set is used as the test set. The experimental results are shown in **Table 4**.

3.3.2. Experiment 3.2: Two-Level Ensemble: Multi-Class Classification

In this experiment, we directly apply an XGBoost classifier on the probability maps from three DeepMedics and three 3D U-Nets. The input vector of the XGBoost classifier has 10 dimensions (5-class probability maps from 2 ensembles of the same type of models). The XGBoost classifier outputs the 5-class labels which contain a background (label 0), enhancing tumor (label 1), edema (label 2), and necrosis & non-enhancing tumor (label 4). BraTS 2018 training set is used to train the models with five-fold cross-validation, and the BraTS 2018 validation set is used as the test set. The experimental results are shown in **Table 4** as TLMC.

3.3.3. Experiment 3.3: Two-Level Ensemble: Binary Classification

In this experiment, we train three XGBoost binary classifiers on the resulting probability maps generated from three DeepMedics and three 3D U-Nets in the first level. During training, each classifier uses a one-vs.-one approach to distinguish between two binary classes. We trained three different models namely, model_WT (whole tumor), model_TC (tumor core), and model_ET (enhancing tumor) as shown in **Figure 6**.

$$\begin{aligned}
&\text{For model_WT:} \quad C_+ = C_{WT} \quad \text{and} \quad C_- = C_{background} \\
&\text{For model_TC:} \quad C_+ = C_{TC} \quad \text{and} \quad C_- = C_{WT} \\
&\text{For model_ET:} \quad C_+ = C_{ET} \quad \text{and} \quad C_- = C_{TC}
\end{aligned}$$

The whole tumor region is the union of edema, non-enhancing tumor & necrosis, and enhancing tumor, and the tumor core regions is the union of edema and non-enhancing tumor & necrosis. Therefore, the tumor core class is a subset class of the whole tumor, and the enhancing tumor class is a subset of the tumor core class. For prediction, we feed the average probability maps from three DeepMedics and three 3D U-Nets

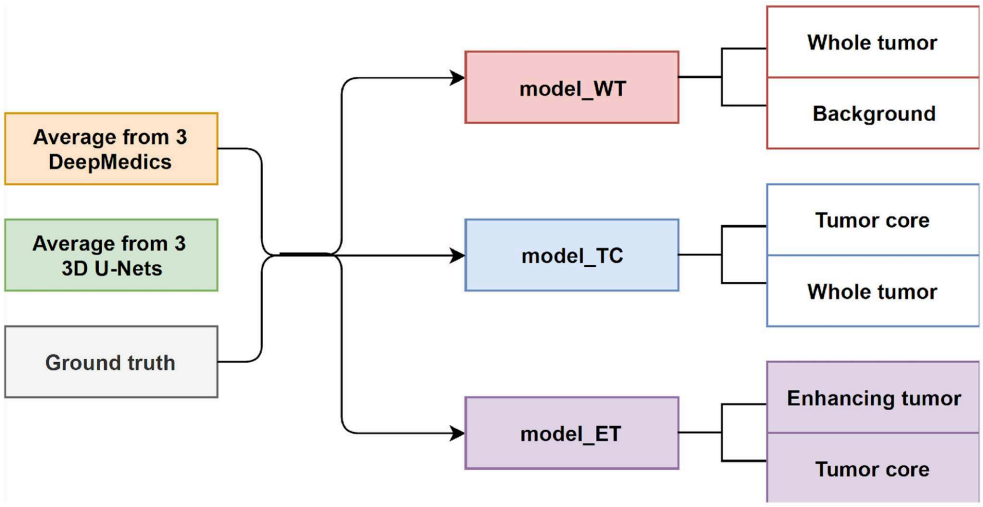


FIGURE 6 | The training workflow of two-level binary classification approach.

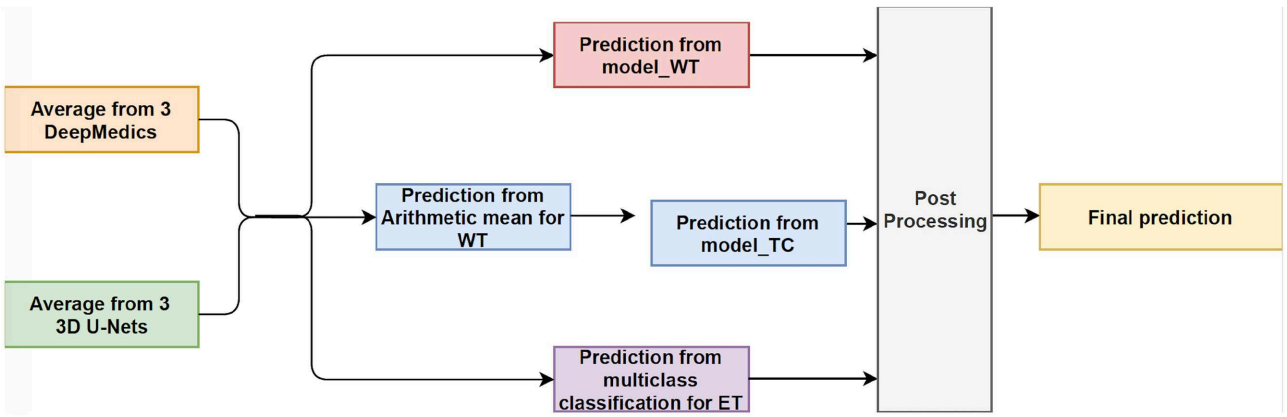


FIGURE 7 | The workflow of Fusion Classification method. For post processing step, we classify the voxels into three classes with background, whole tumor, tumor core, and necrosis & non-enhancing tumor in decreasing order of priority. For example, if a voxel is classified into both whole tumor and tumor core, we give the final label as that of tumor core according to the preference mentioned before.

to the three models. The input vector has 10 dimensions (5-class probability maps from 2 ensembles of the same type of models). The model_WT classifies the voxels into the whole tumor and background. For model_TC, we feed the probability maps of such voxels that are classified as the whole tumor from the experiment in section 3.3.1. For model_ET, we feed the probability maps of such voxels that are classified as tumor core from the previous prediction in the experiment. BraTS 2018 training set is used to train the models with five-fold cross-validation, and the BraTS 2018 validation set is used as the test set. The experimental results are shown in **Table 4** as TLBC.

3.3.4. Experiment 3.4: Two-Level Ensemble: Fusion Classifications

This is the final experiment to integrate the methods from the previous experiments. We observe that while the experiment

TABLE 5 | The first three rows show the results of our proposed method and the state-of-the-art methods on the BraTS 2017 validation set, and the bottom four rows show the results of our proposed method and the state-of-the-art methods on BraTS 2018 validation set.

Methods	No. of models	DSC			HD95		
		ET	WT	TC	ET	WT	TC
Kamnitsas et al. (2017a)	7	73.8	90.1	**79.7**	4.50	**4.23**	**6.56**
Isensee et al. (2017)	5	73.2	89.6	**79.7**	4.55	6.97	9.48
Proposed method	6	**74.3**	**90.4**	78.5	**3.49**	4.46	8.45
Myronenko (2018)	10	**82.3**	**91.0**	**86.6**	3.93	4.52	**6.85**
Isensee et al. (2018)	10	81.0	90.8	85.4	**2.54**	4.97	7.04
Kao et al. (2018)	26	78.8	90.5	81.3	3.81	**4.32**	7.56
Proposed method	6	78.2	90.8	82.3	2.96	4.39	6.91

The results are reported as mean. Bold numbers highlight the best performance in each dataset. These results are directly copied from their paper.

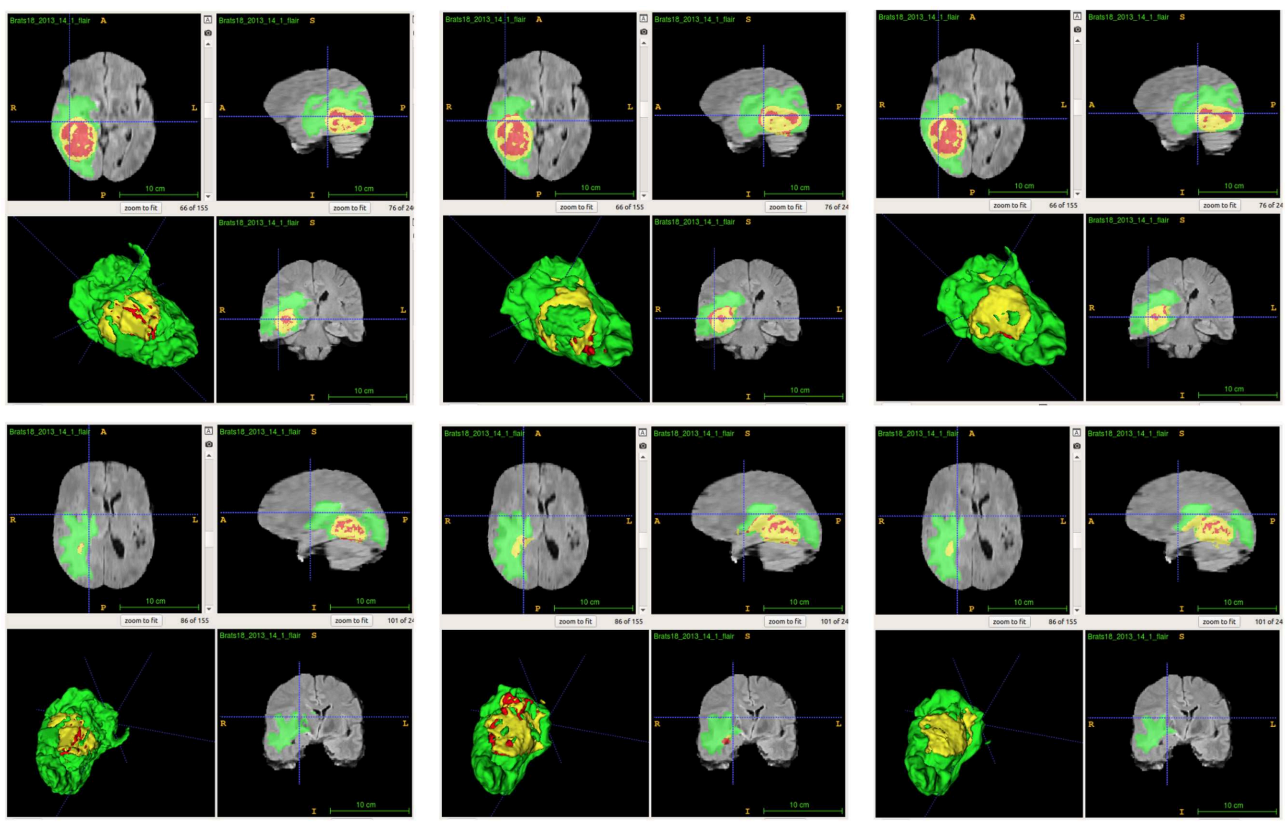


FIGURE 8 | Examples of predictions from single model with different inputs. Top row shows the predictions from DeepMedic, and bottom row shows the predictions from 3D U-Net (from left to right: ground-truth lesions, prediction from single model, and prediction from single model with additional brain parcellation masks.) Red: enhancing tumor, yellow: necrosis & non-enhancing tumor, and green: edema. ITK-SNAP (Fedorov et al., 2012) is used to visualize the MR images and lesion masks.

in section 3.3.3 performs best for classifying voxels into the background, whole tumor and tumor core, the experiment in section 3.3.1 has the best performance on necrosis & non-enhancing tumor. We use model_WT, model_TC, and multi-class classifier model for the fusion model. For prediction, we feed the average probability maps from three DeepMedics and three 3D U-Nets to the three models. The input vector has 10 dimensions (5-class probability maps from 2 ensembles of the same type of models). The model_WT classifies the voxels into the whole tumor and background. For model_TC that is trained to classify voxels into the whole tumor and tumor core, we feed the probability maps of such voxels that are classified as the whole tumor from the experiment in section 3.3.1. For necrosis & non-enhancing tumor class, we feed the probability maps to the multi-class classifier as in section 3.3.2. To merge the three different predicted results, we classify the voxels into three classes with background, whole tumor, tumor core, and necrosis & non-enhancing tumor in decreasing order of priority. For example, if a voxel is classified into both whole tumor and tumor core, we give the final label as that of tumor core according to the preference mentioned before. Therefore integrating these two gives the effective scores as shown in **Table 4** as TLFC. BraTS 2018 training set is used to train the models with five-fold cross-validation, and the BraTS 2018 validation set is used as the test set. The workflow of fusion classification is shown in **Figure 7**.

3.4. Experiment 4: Compare to the State-of-the-Art Methods

In this experiment, we compare the brain tumor segmentation performance of the proposed method described in section 3.3.4 with the state-of-the-art methods on both BraTS 2017 and BraTS 2018 dataset. The quantitative results are shown in **Table 5**.

4. DISCUSSION AND CONCLUSION

Due to the computational limitation of training the state-of-the-art networks using GPU, we are not able to input the whole brain volume of size 240 × 240 × 155 to a neural network for training purposes. Alternatively, we randomly crop sub-regions of the brain and input these sub-regions to the neural network for training. For the current patch-based neural networks, we noted that these neural networks lack location information of the brain for both training and test procedure. That is, these patch-based neural networks do not have the information about where the patch comes

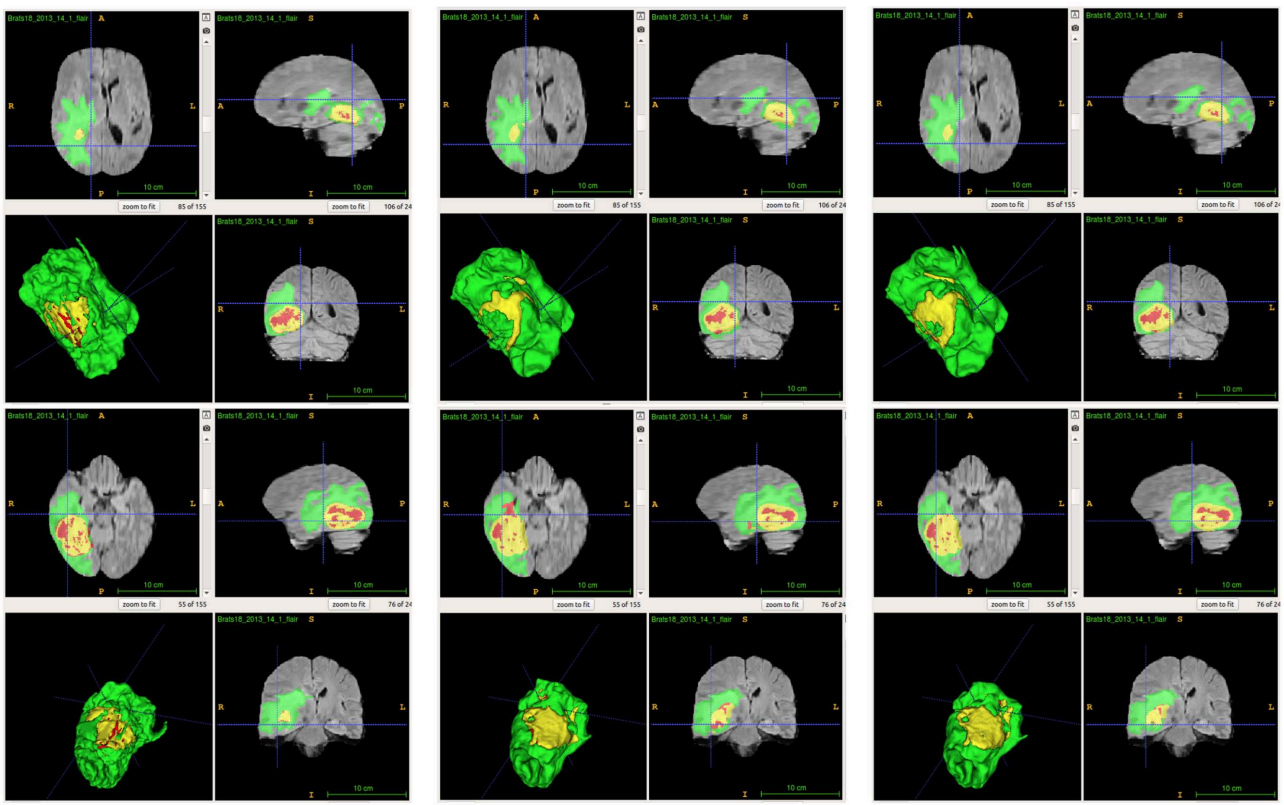


FIGURE 9 | Examples of predictions from ensemble with different inputs. Top row shows the predictions from ensemble of DeepMedics, and bottom row shows the predictions from ensemble of 3D U-Nets (from left to right: ground-truth lesions, prediction from ensemble, and prediction from ensemble with additional brain parcellation masks.) Red: enhancing tumor, yellow: necrosis & non-enhancing tumor, and green: edema. ITK-SNAP (Fedorov et al., 2012) is used to visualize the MR images and lesion masks.

from the brain. Therefore, we proposed the location fusion method which explicitly carries location information of the brain into patch-based neural networks such as 3D U-Net and DeepMedic. An existing structural brain parcellation atlas, HarvardOxford Sub-cortical Atlas, is used as additional location information to these patch-based neural networks in both training and test.

From **Table 2**, we demonstrate that the proposed location fusion method improves the brain tumor segmentation performance of both a single state-of-the-art model. We also demonstrate that the proposed location fusion method improves the ensemble of multiple same types of state-of-the-art models in **Table 3**. The proposed location fusion method yields a smoother prediction for both 3D U-Net and DeepMedic compared to the resulting prediction without location information (see **Figures 8**, **9**).

From **Table 4**, the proposed ensemble method, two-level fusion classification (TLFC) method, has the best performance compared to other ensemble methods including arithmetic mean, two-level multi-class classification (TLMC), and two-level binary classification (TLBC). TLFC takes advantage of TLMC and TLBC. Moreover, **Figure 10** shows the predictions of brain tumor lesions from different ensemble methods, and TLFC method has the best performance among other methods.

From **Table 5**, the proposed method has the best tumor segmentation performance compared to other state-of-the-art methods in BraTS 2017 with a similar number of models in the ensemble. Also, the proposed method has a competitive tumor segmentation performance compared to other state-of-the-art methods in BraTS 2018 with fewer models in the ensemble. It is noted that the model of Myronenko (2018) requires a large amount of GPU memory (32 GB) for training,

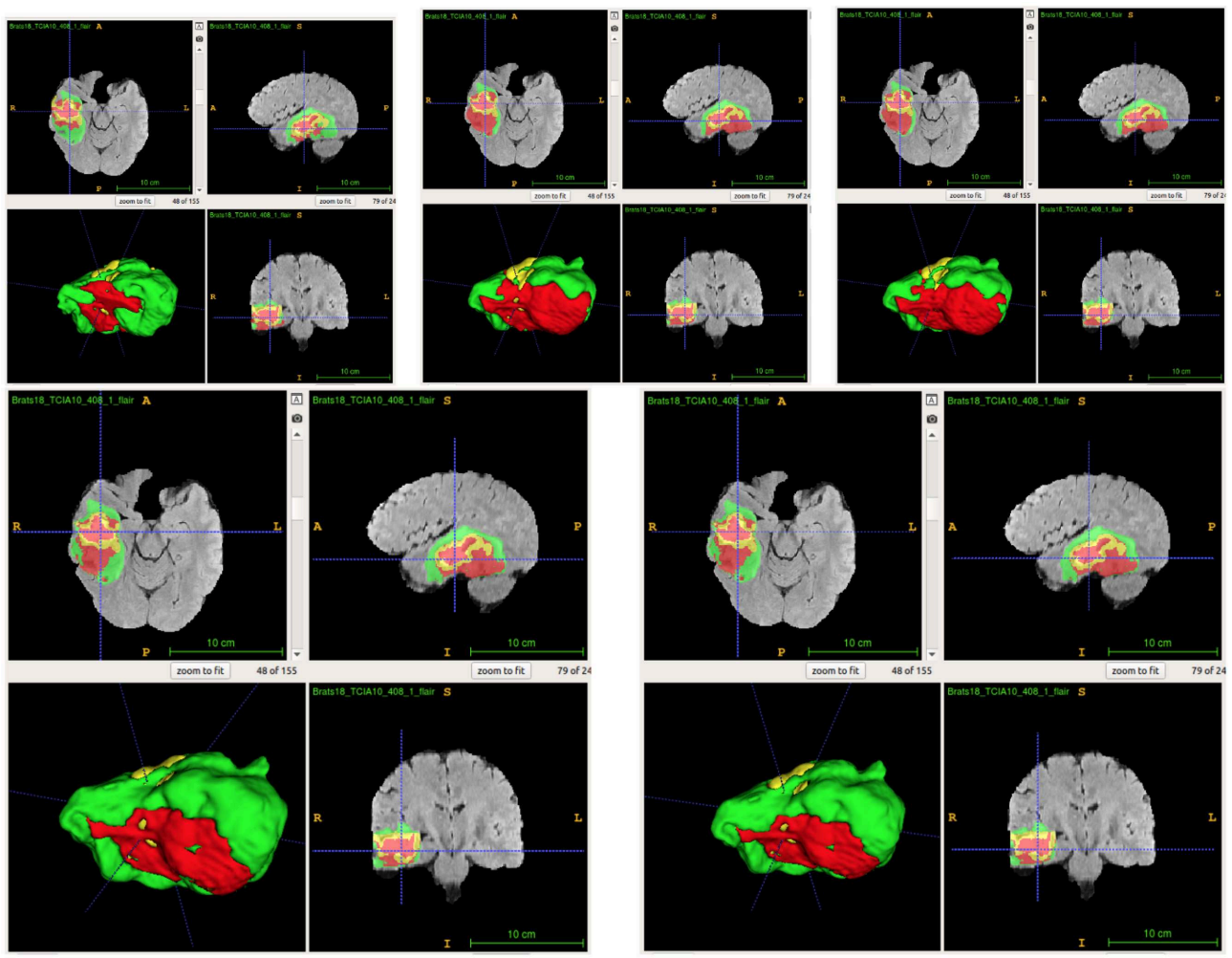


FIGURE 10 | Examples of predictions from different ensemble methods. The top left image shows the ground-truth lesion mask, and the top middle image shows the predictions using the arithmetic mean. The top right image shows the prediction using a two-level multi-class classification (TLMC) method. The bottom left image shows the prediction using a two-level binary classification (TLBC) method, and the bottom right image shows the prediction using a two-level fusion classification (TLFC) method. Red: enhancing tumor, yellow: necrosis & non-enhancing tumor, and green: edema. ITK-SNAP (Fedorov et al., 2012) is used to visualize the MR images and lesion masks.

and Isensee et al. (2018) trained the models with additional public and institutional data. In addition, Myronenko (2018) and Isensee et al. (2018) have 10 models in their ensemble but our proposed ensemble only has six models. The proposed ensemble has much fewer models with a better segmentation performance compared to our previous work which has 26 models (Kao et al., 2018). The test time of our previous ensemble takes approximately 30 min on an Nvidia 1080 Ti GPU and an Intel Xeon CPU E5-2696 v4 @ 2.20 GHz. However, the proposed ensemble only takes approximate 3 min on the same infrastructure. Our previous ensemble ranked 6th out of 63 teams in BraTS 2018 segmentation challenge, and the proposed ensemble even has a better performance and less inference time compared to the previous ensemble.

Summarizing, in this paper we proposed a novel method to integrate location information about the brain into a patch-based neural network for improving brain tumor segmentation. Our experimental results demonstrate that the proposed location information fusion approach improves the segmentation performance of the baseline models including DeepMedic and 3D U-Net. Moreover, the proposed location information fusion method can be easily integrated with other patch-based network architectures to potentially enhance their brain tumor segmentation performance. We also proposed a two-level fusion classification method which reduces the uncertainty of prediction in the first level and takes advantage of different types of models in the second level. Also, the proposed ensemble method can also be easily integrated with more different types of neural networks. The proposed ensemble helps the neurologists on delineating brain tumors and improves the quality of the neuro-surgery.

DATA AVAILABILITY STATEMENT

The dataset analyzed for this study can be found in the BraTS 2017 and BraTS 2018 of CBICA Image Processing Portal [https://ipp.cbica.upenn.edu/].

AUTHOR'S NOTE

All data are made available online with a permissive copyright license (CC-BY-SA 4.0), allowing for data to be shared, distributed, and improved upon.

AUTHOR CONTRIBUTIONS

P-YK, SS, and JJ designed the study and algorithm with BM guiding the research and analyzed and interpreted the data. P-YK collected the data. P-YK and SS sourced the literature and wrote the draft. P-YK, SS, JJ, AZ, AK, JC, and BM edited the manuscript. JC provided medical and clinical insights.

FUNDING

This research was partially supported by a National Institutes of Health (NIH) award # 5R01NS103774-03.

ACKNOWLEDGMENTS

We thank Oytun Ulutan for technical support and Dr. Robby Nadler for writing assistance and language editing.

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). *Segmentation Labels and Radiomic Features for the Pre-Operative Scans of the TCGA-GBM Collection*. Technical Report.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). *Segmentation Labels and Radiomic Features for the Pre-Operative Scans of the TCGA-LGG Collection.*

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). 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*.

Bühlmann, P. (2012). "Bagging, boosting and ensemble methods," in *Handbook of Computational Statistics* (Springer), 985–1022. Available online at: https://www.econstor.eu/bitstream/10419/22204/1/31_pb.pdf

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM), 785–794.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 424–432.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Dietterich, T. G. (2000). "Ensemble methods in machine learning," in *International Workshop on Multiple Classifier Systems* (Springer), 1–15.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., et al. (2012). 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* 30, 1323–1341. doi: 10.1016/j.mri.2012.05.001

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi: 10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., Collins, D. L. et al. (2006). "Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 58–66.

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. *Stat. Interf.* 2, 349–360.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2017). "Brain tumor segmentation and radiomics survival prediction:

contribution to the brats 2017 challenge," in *International MICCAI Brainlesion Workshop* (Springer), 287–297.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Springer), 234–244.

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156. doi: 10.1016/s1361-8415(01)00036-6

Jiang, J., Kao, P.-Y., Belteton, S. A., Szymanski, D. B., and Manjunath, B. (2019). Accurate 3D cell segmentation using deep feature and CRF refinement. *arXiv preprint arXiv:1902.04729*. doi: 10.1109/ICIP.2019.8803095

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2017a). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI Brainlesion Workshop* (Springer), 450–462.

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., and Glocker, B. (2015). Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. *Ischem. Stroke Lesion Segment.* 13, 13–16.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., et al. (2016). "Deepmedic for brain tumor segmentation," in *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Springer), 138–149.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017b). Efficient multi-scale 3D CNN with fully connected crf for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kao, P.-Y., Ngo, T., Zhang, A., Chen, J. W., and Manjunath, B. (2018). "Brain tumor segmentation and tractographic feature extraction from structural MR images for overall survival prediction," in *International MICCAI Brainlesion Workshop* (Springer), 128–141.

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations*.

Kuijf, H. J., Biesbroek, J. M., Viergever, M. A., Biessels, G. J., and Vincken, K. L. (2013). "Registration of brain CT images to an MRI template for the purpose of lesion-symptom mapping," in *International Workshop on Multimodal Brain Image Analysis*, (Springer), 119–128.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, 6402–6413.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* 37, 2663–2674. doi: 10.1109/TMI.2018.2845918

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., et al. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* 114, 97–109. doi: 10.1007/s00401-007-0243-4

McKinley, R., Meier, R., and Wiest, R. (2018). "Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Springer), 456–465.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Springer), 311–320.

Nielsen, D. (2016). *Tree boosting with XGBoost-why does XGBoost win "every" machine learning competition?* Master's thesis, NTNU.

Noiphithak, R., and Veerasarn, K. (2017). Clinical predictors for survival and treatment outcome of high-grade glioma in prasat neurological institute. *Asian J. Neurosurg.* 12, 28–33. doi: 10.4103/1793-5482.148791

Reddi, S. J., Kale, S., and Kumar, S. (2018). "On the convergence of adam and beyond," in *International Conference on Learning Representations*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 234–241.

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Springer), 178–190.

Wu, Y., and He, K. (2018). "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Yu, L., Yang, X., Chen, H., Qin, J., and Heng, P. A. (2017). "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhou, C., Chen, S., Ding, C., and Tao, D. (2018). "Learning contextual and attentive information for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Springer), 497–507.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

CORRECTION
published: 15 April 2020
doi: 10.3389/fnins.2020.00328

Corrigendum: Improving Patch-Based Convolutional Neural Networks for MRI Brain Tumor Segmentation by Leveraging Location Information

Po-Yu Kao[1], Shailja Shailja[1], Jiaxiang Jiang[1], Angela Zhang[1], Amil Khan[1], Jefferson W. Chen[2] and B. S. Manjunath[1]**

[1] Vision Research Lab, Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA, United States, [2] Department of Neurological Surgery, University of California, Irvine, Irvine, CA, United States

Keywords: gliomas, brain tumor segmentation, brain parcellation atlas, convolutional neural network, DeepMedic, 3D U-Net, ensemble learning, XGBoost

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

****Correspondence:***
Po-Yu Kao
poyu_kao@ucsb.edu
B. S. Manjunath
manj@ucsb.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *10 March 2020*
Accepted: *19 March 2020*
Published: *15 April 2020*

Citation:
Kao P-Y, Shailja S, Jiang J, Zhang A, Khan A, Chen JW and Manjunath BS (2020) Corrigendum: Improving Patch-Based Convolutional Neural Networks for MRI Brain Tumor Segmentation by Leveraging Location Information. Front. Neurosci. 14:328. doi: 10.3389/fnins.2020.00328

A Corrigendum on

Improving Patch-Based Convolutional Neural Networks for MRI Brain Tumor Segmentation by Leveraging Location Information

by Kao, P.-Y., Shailja, S., Jiang, J., Zhang, A., Khan, A., Chen, J. W., et al. (2020). Front. Neurosci. 13:1449. doi: 10.3389/fnins.2019.01449

An author's name was incorrectly published as "**Fnu Shailja.**" It should be "**Shailja Shailja.**" The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 07 February 2020
doi: 10.3389/fncom.2020.00006

Demystifying Brain Tumor Segmentation Networks: Interpretability and Uncertainty Analysis

*Parth Natekar, Avinash Kori and Ganapathy Krishnamurthi**

Department of Engineering Design, Indian Institute of Technology Madras, Chennai, India

The accurate automatic segmentation of gliomas and its intra-tumoral structures is important not only for treatment planning but also for follow-up evaluations. Several methods based on 2D and 3D Deep Neural Networks (DNN) have been developed to segment brain tumors and to classify different categories of tumors from different MRI modalities. However, these networks are often black-box models and do not provide any evidence regarding the process they take to perform this task. Increasing transparency and interpretability of such deep learning techniques is necessary for the complete integration of such methods into medical practice. In this paper, we explore various techniques to explain the functional organization of brain tumor segmentation models and to extract visualizations of internal concepts to understand how these networks achieve highly accurate tumor segmentations. We use the BraTS 2018 dataset to train three different networks with standard architectures and outline similarities and differences in the process that these networks take to segment brain tumors. We show that brain tumor segmentation networks learn certain human-understandable disentangled concepts on a filter level. We also show that they take a top-down or hierarchical approach to localizing the different parts of the tumor. We then extract visualizations of some internal feature maps and also provide a measure of uncertainty with regards to the outputs of the models to give additional qualitative evidence about the predictions of these networks. We believe that the emergence of such human-understandable organization and concepts might aid in the acceptance and integration of such methods in medical diagnosis.

Keywords: interpretability, CNN, brain tumor, segmentation, uncertainty, activation maps, features, explainability

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Fan Zhang,
Harvard Medical School,
United States
Hongming Li,
University of Pennsylvania,
United States

****Correspondence:***
Ganapathy Krishnamurthi
gankrish@iitm.ac.in

Received: *07 September 2019*
Accepted: *17 January 2020*
Published: *07 February 2020*

Citation:
Natekar P, Kori A and Krishnamurthi G (2020) Demystifying Brain Tumor Segmentation Networks: Interpretability and Uncertainty Analysis. Front. Comput. Neurosci. 14:6. doi: 10.3389/fncom.2020.00006

1. INTRODUCTION

Deep learning algorithms have shown great practical success in various tasks involving image, text and speech data. As deep learning techniques start making autonomous decisions in areas like medicine and public policy, there is a need to explain the decisions of these models so that we can understand *why* a particular decision was made (Molnar, 2018).

In the field of medical imaging and diagnosis, deep learning has achieved human-like results on many problems (Esteva et al., 2017; Weng et al., 2017; Kermany et al., 2018). Interpreting the decisions of such models in the medical domain is especially important, where transparency and a clearer understanding of Artificial Intelligence are essential from a regulatory point of view and to make sure that medical professionals can trust the predictions of such algorithms.

Understanding the organization and knowledge extraction process of deep learning models is thus important. Deep neural networks often work in higher dimensional abstract concepts. Reducing these to a domain that human experts can understand is necessary—if a model represents the underlying data distribution in a manner that human beings can comprehend and a logical hierarchy of steps is observed, this would provide some backing for its predictions and would aid in its acceptance by medical professionals.

However, while there has been a wide range of research on Explainable AI in general (Doshi-Velez and Kim, 2017; Gilpin et al., 2018), it has not been properly explored in the context of deep learning for medical imaging. Holzinger et al. (2017) discuss the importance of interpretability in the medical domain and provide an overview of some of the techniques that could be used for explaining models which use the image, omics, and text data.

In this work, we attempt to extract explanations for models which accurately segment brain tumors, so that some evidence can be provided regarding the process they take and how they organize themselves internally. We first discuss what interpretability means with respect to brain tumor models. We then present the results of our experiments and discuss what these could imply for machine learning assisted tumor diagnosis.

2. INTERPRETABILITY IN THE CONTEXT OF BRAIN TUMOR SEGMENTATION MODELS

Interpreting deep networks which accurately segment brain tumors is important from the perspectives of both transparency and functional understanding (by functional understanding, we mean understanding the role of each component or filter of the network and how these relate to each other). Providing glimpses into the internals of such a network to provide a *trace of its inference steps* (Holzinger et al., 2017) would go at least some way to elucidating exactly how the network makes its decisions, providing a measure of legitimacy.

There have been several methods explored for trying to look inside a deep neural network. Many of these focus on visual interpretability, i.e., trying to extract understandable visualizations from the inner layers of the network or understanding what the network looks at when giving a particular output (Zhang and Zhu, 2018).

For a brain tumor segmentation model, such methods might provide details on how information flows through the model and how the model is organized. For example, it might help in understanding how the model represents information regarding the brain and tumor regions internally, and how these representations change over layers. Meaningful visualizations of the internals of a network will not only help medical professionals in assessing the legitimacy of the predictions but also help deep learning researchers to debug and improve performance.

In this paper, we aim to apply visual interpretability and uncertainty estimation techniques on a set of models with different architectures to provide human-understandable visual interpretations of some of the concepts learned by different parts of a network and to understand more about the organization of these different networks. We organize our paper into mainly three parts as described in **Figure 1**: (1) Understanding information organization in the model, (2) Extracting visual representations of internal concepts, and (3) Quantifying uncertainty in the outputs of the model. We implement our pipeline on three different 2D brain tumor segmentation models—a Unet model with a densenet121 encoder (Henceforth referred to as the DenseUnet) (Shaikh et al., 2017), a Unet model with a ResNet encoder (ResUnet) (Kermi et al., 2018), and a simple encoder-decoder network which has a similar architecture to the ResUnet but without skip or residual connections (SimUnet). All models were trained till convergence on the BraTS

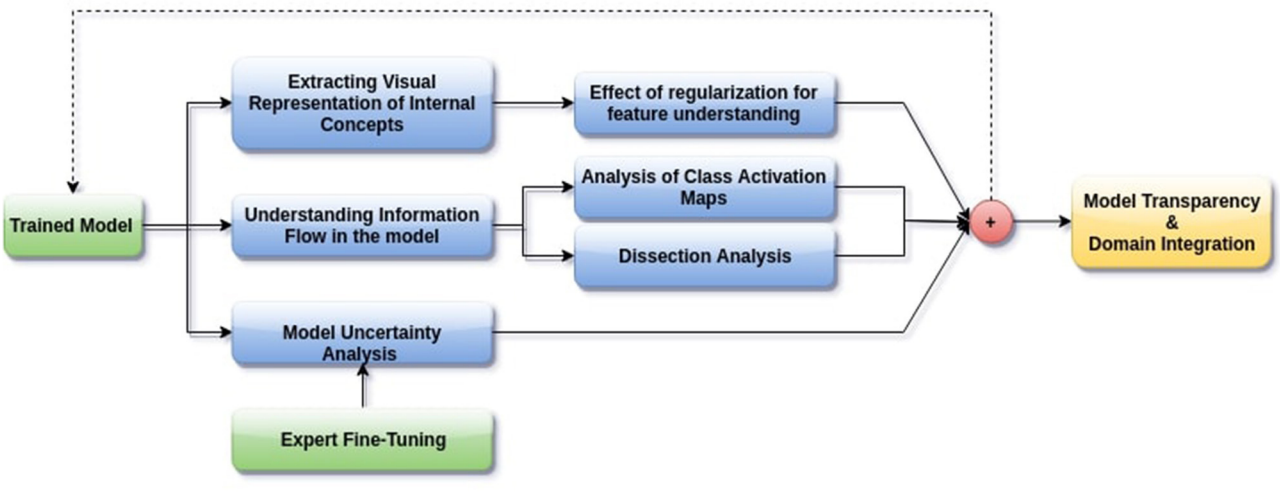


FIGURE 1 | Proposed pipeline for interpreting brain tumor segmentation models to aid in increasing transparency. The dotted backward arrow shows the possibility of using the inferences from such an experiment to enhance the training process of networks.

TABLE 1 | Performance metrics of our networks.

Model type	WT dice	TC dice	ET dice
DenseUnet	0.830	0.760	0.685
ResUnet	0.788	0.734	0.649
SimUnet	0.743	0.693	0.523

WT, Whole Tumor; TC, Tumor Core; ET, Enhancing Tumor.

2018 dataset (Menze et al., 2014; Bakas et al., 2017a,b, 2018). A held out validation set of 48 volumes (including both LGG and HGG volumes) was used for testing. **Table 1** shows the performance of the three models on this test set.

Our models are not meant to achieve state of the art performance. Instead, we aim to demonstrate our methods on a set of models with different structures commonly used for brain tumor segmentation and compare them to better understand the process they take to segment the tumors. In this primary study, we do not use 3D models, since the visualization and analysis of interpretability related metrics is simpler for 2D models. Also, it is not clear how some of our results would scale to 3D models and whether it would be possible to visualize these. For example, disentangled concepts observed by performing network dissection might not be meaningful when visualized slice wise and would have to be visualized in 3D. This and the related analysis poses an additional layer of difficulty.

We now give a brief introduction of each interpretability techniques in our pipeline. *Network Dissection* aims to quantify to what extent internal information representation in CNNs is human interpretable. This is important to understand what concepts the CNN is learning on a filter level, and whether these correspond with human level concepts. *Grad-CAM* allows us to see how the spatial attention of the network changes over layers, i.e., what each layer of the network looks at in a specific input image. This is done by finding the importance of each neuron in the network by taking the gradient of the output with respect to that neuron. In *feature visualization*, we find the input image which maximally activates a particular filter, by randomly initializing an input image and optimizing this for a fixed number of iterations, referred to as *activation maximization*. Such an optimized image is assumed to be a good first order representation of the filter, which might allow us to understand how a neural network "sees." *Test-time dropout* is a computationally efficient method of approximate Bayesian Inference on a CNN to quantify uncertainty in the outputs of the model.

In the following sections, each element of the proposed pipeline is implemented and its results and implications are discussed.

3. UNDERSTANDING INFORMATION ORGANIZATION IN THE MODEL

3.1. Network Dissection

Deep neural networks may be learning explicit disentangled concepts from the underlying data distribution. For example, Zhou et al. (2014) show that object detectors emerge in networks trained for scene classification. To study whether filters in brain tumor segmentation networks learn such disentangled concepts, and to quantify such functional disentanglement (i.e., to quantify to what extent individual filters learn individual concepts), we implement the Network Dissection (Bau et al., 2017) pipeline, allowing us to determine the function of individual filters in the network.

In-Network Dissection, the activation map of an internal filter for every input image is obtained. Then the distribution α of the activation is formulated over the entire dataset. The obtained activation map is then resized to the dimensions of the original image and thresholded to get a concept mask. This concept mask might tell us which individual concept a particular filter learns when overlaid over the input image.

For example, in the context of brain-tumor segmentation, if the model is learning disentangled concepts, there might be separate filters learning to detect, say, the edema region, or the necrotic tumor region. The other possibility is that the network somehow spreads information in a form not understandable by humans - entangled and non-interpretable concepts.

Mathematically, Network Dissection is implemented by obtaining activation maps $\Phi_{k,l}$ of a filter k in layer l, and then obtaining the pixel level distribution α of $\Phi_{k,l}$ over the entire dataset.

A threshold $T_{k,l}(x)$ is determined as the 0.01-quantile level of $\alpha_{k,l}(x)$, which means only 1.0% of values in $\Phi_{k,l}(x)$ are greater than $T_{k,l}(x)$. (We choose the 0.01-quantile level since this gives the best results qualitatively (visually) and also quantitatively in terms of dice score for the concepts for which ground truths are available). The concept mask is obtained as:

$$M_{k,l}(x) = \Phi_{k,l}(x) \geq T_{k,l}(x) \tag{1}$$

A channel is a detector for a particular concept if:

$$IoU(M_{k,l}(x), gt) = \frac{|M_{k,l}(x) \cap gt|}{|M_{k,l}(x) \cup gt|} \geq c \tag{2}$$

In this study, we only quantify explicit concepts like the core and enhancing tumor due to the availability of ground truths gt and recognize detectors for other concepts by visual inspection. We post-process the obtained concept images to remove salt-and-pepper noise and keep only the largest activated continuous concept inside the brain region in the image. The IoU between the final concept image and the ground truth for explicit concepts is used to determine the quality of the concept.

The results of this experiment, shown in **Figures 2–4**, indicate that individual filters of brain-tumor segmentation networks learn explicit as well as implicit disentangled concepts. For example, **Figure 2E** shows a filter learning the concept *whole tumor region* i.e., it specifically detects the whole tumor region for any image in the input distribution, the filter in **Figure 2B** seems to be learning the *edema region*, while **Figure 2A** shows a filter learning the *white and gray matter region*, an implicit concept which the network is not trained to learn. Similar behavior is seen in all networks (**Figures 2–4**). This means that we can make

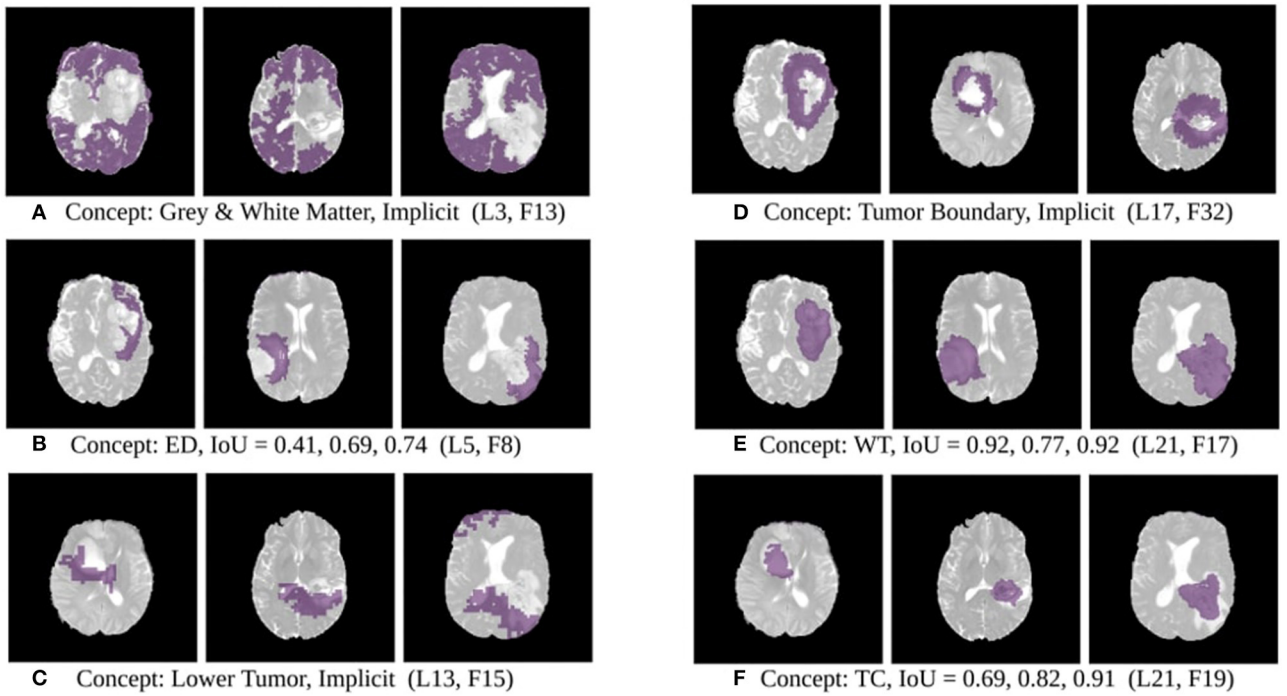


FIGURE 2 | Disentangled concept mask *M* learned by individual filters of the ResUnet overlaid over brain image. This includes explicit concepts for which ground truth labels are available as well as implicit concepts for which their are no labels. IoU scores are mentioned in the sub-captions for all 3 images. L, Layer; WT, Whole Tumor; TC, Tumor Core; ED, Edema.

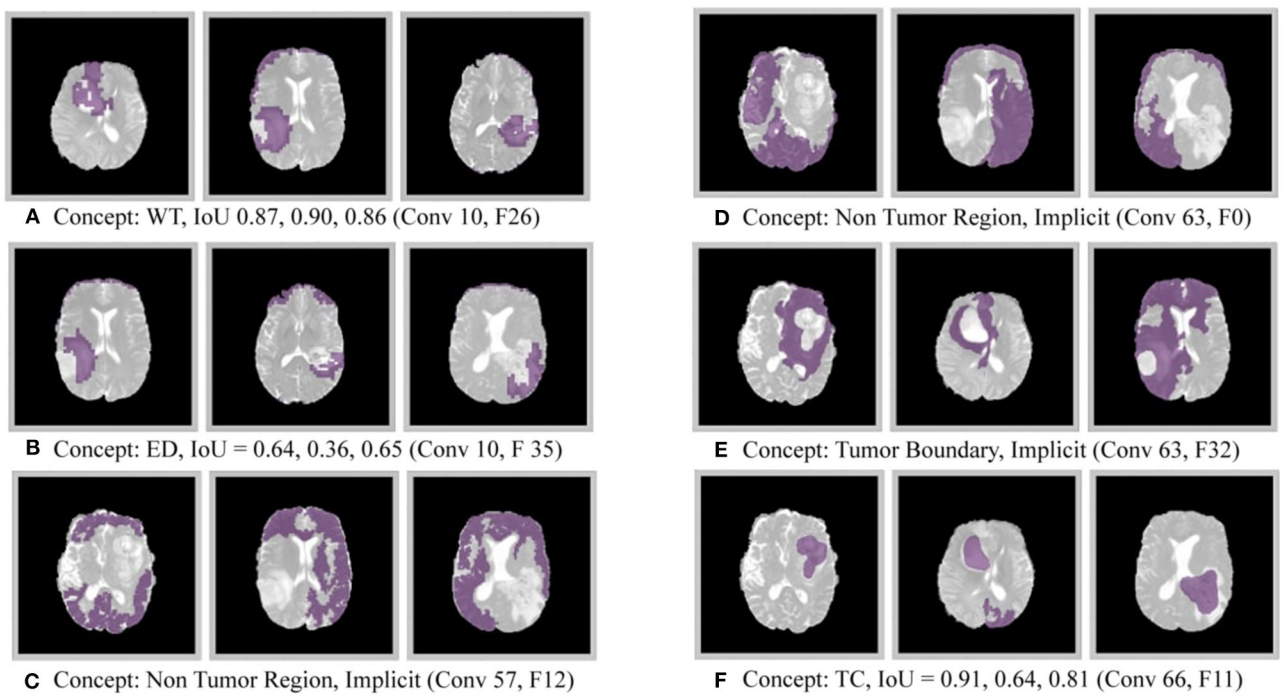


FIGURE 3 | Disentangled concepts learned by filters of the DenseUnet. L, Layer; WT, Whole Tumor; TC, Tumor Core; ED: Edema.

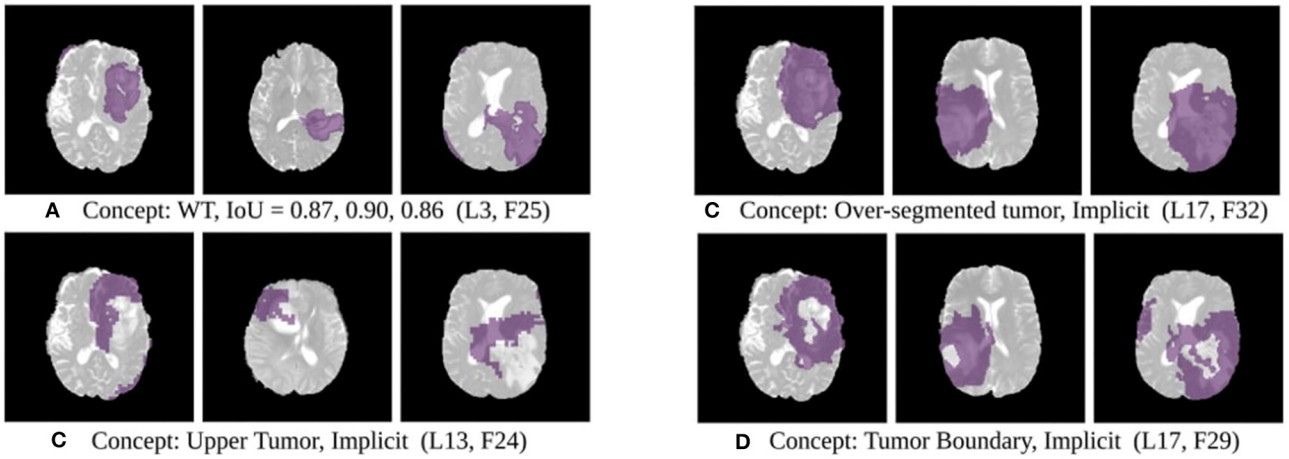


FIGURE 4 | Disentangled concepts learned by filters of the SimUnet. L, Layer; WT, Whole Tumor; TC, Tumor Core; ED: Edema.

attributions based on function to the network at a filter level—indicating a sort of functional specificity in the network i.e., individual filters might be specialized to learn separate concepts.

Neural Networks are inspired by neuroscientific principles. What does this functional specificity mean in this context? Debates are ongoing on whether specific visual and cognitive functions in the brain are segregated and the degree to which they are independent. Zeki and Bartels (1998) discuss the presence of spatially distributed, parallel processing systems in the brain, each with its separate function. Neuroscientific studies have shown that the human brain has some regions that respond specifically to certain concepts, like the face fusiform area Kanwisher and Yovel (2006)—indicating certain visual modularity. Studies based on transcranial magnetic stimulation of the brain also show separate areas of the visual cortex play a role in detecting concepts like faces, bodies, and objects (Pitcher et al., 2009).

The emergence of concept detectors in our study indicates that brain-tumor segmentation networks might show a similar modularity. This indicates that there is some organization in the model similar to the process a human being might take to recognize a tumor, which might have an implications with regards to the credibility of these models in the medical domain, in the sense that they might be taking human-like, or at least human understandable, steps for inference.

The extracted disentangled concepts can also be used for providing contextual or anatomical information as feedback to the network. Though we do not explore this in this study, 3D concept maps obtained from networks can be fed back as multi-channel inputs to the network to help the network implicitly learn to identify anatomical regions like the gray and white matter, tumor boundary etc. for which no labels are provided, which might improve performance. This would be somewhat similar to the idea of feedback networks discussed by Zamir et al. (2017), where an implicit taxonomy or hierarchy can be established during training as the network uses previously learned concepts to learn better representations and increase speed of learning.

3.2. Gradient Weighted Class Activation Maps

Understanding how spatial attention of a network over an input image develops might provide clues about the overall strategy the network uses to localize and segment an object. Gradient weighted Class Activation Maps (Grad-CAM) (Selvaraju et al., 2017) is one efficient technique that allows us to see the networks attention over the input image. Grad-CAM provides the region of interest on an input image which has a maximum impact on predicting a specific class.

Segmentation is already a localization problem. However, our aim here is to see *how attention changes over internal layers of the network*, to determine how spatial information flows in the model. To understand the attentions of each layer on an input image, we convert segmentation to a multi-label classification problem by considering class wise global average pooling on the final layer. The gradient of the final global average pooled value is considered for attention estimation in Grad-CAM. To understand the layer-wise feature map importance, Grad-CAM was applied to see the attention of every internal layer.

This mathematically amounts to finding neuron importance weights $\beta^c_{l,k}$ for each filter k of a particular layer l with respect to the global average pooled output segmentation for a particular channel c:

$$y(c) = \frac{1}{P}\sum_i\sum_j \Phi^c(x) \tag{3}$$

$$\beta^c_{l,k} = \frac{1}{N}\sum_i\sum_j \frac{\partial y(c)}{\partial A^{ij}_{l,k}(x)} \tag{4}$$

$$O_{GradCAM}(c) = ReLU\left(\sum_k \beta^c_{l,k} A_{l,k}(x)\right) \tag{5}$$

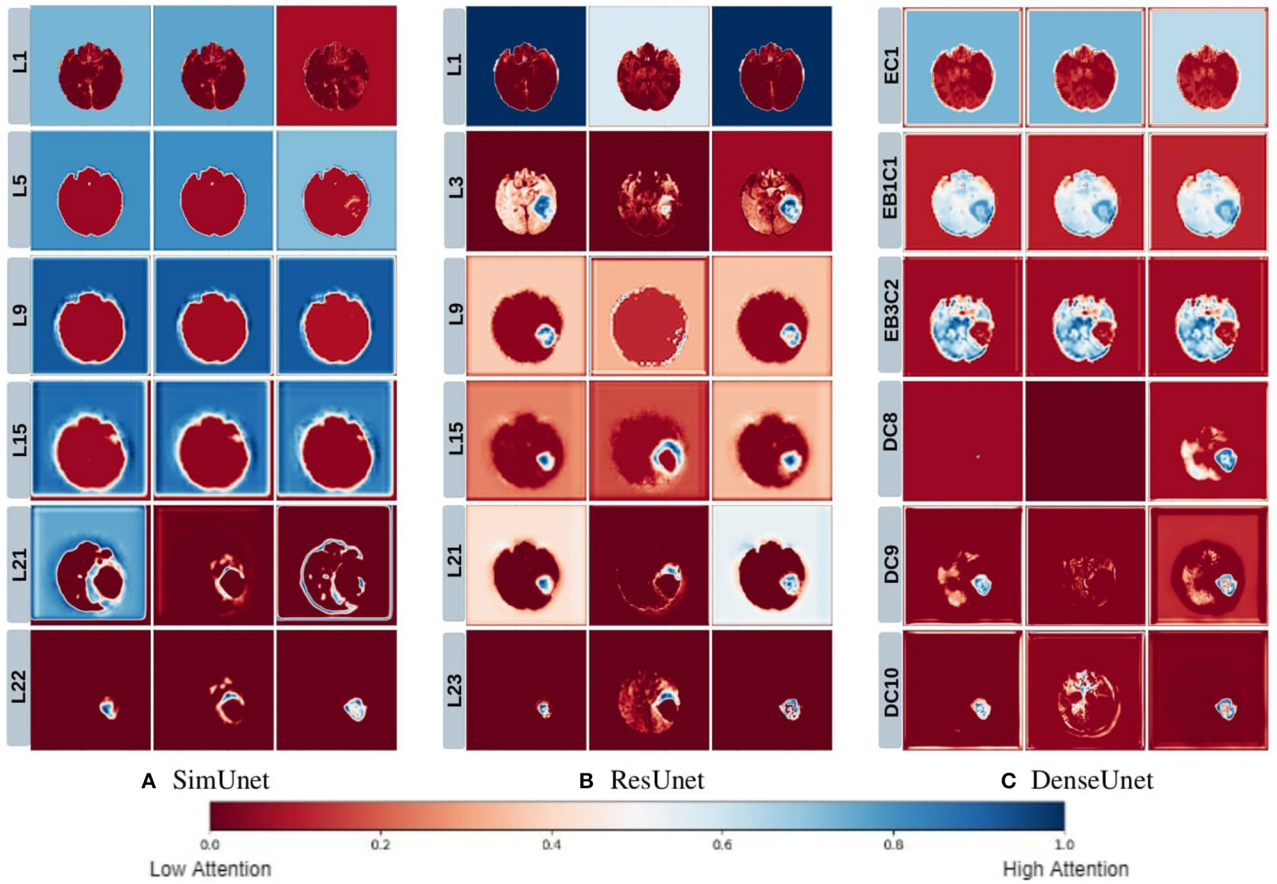


FIGURE 5 | This figure depicts the gradient based class activation maps obtained at selected intermediate layers of all the three networks in ascending order. L, Layer; E, Encoding; B, Block; D, Decoding.

Where, P and N are the number of pixels in the output segmentation map and the activation map of the relevant layer for channel c respectively, Φ^c is the output segmentation map for class c of network Φ, $y(c)$ describes the spatially pooled final segmentation map, $A_{l,k}(x)$ is the activation map for the k^{th} filter of the l^{th} layer, and $O_{GradCAM}(c)$ represents an output map which is the result of *GradCAM* for channel c.

We posit that model complexity and residual connections might have an impact on how early a model can localize the tumor region. For example, the DenseUnet and ResUnet localize the tumor region in the first few layers, while the SimUnet, which has no skip or residual connections, localizes the tumor region only in the final few layers (**Figure 5**). This indicates that skip and residual connections help learn and propagate spatial information to the initial layers for faster localization. While previous literature indicates that skip connections allow upsampling layers to retain fine-grained information from downsampling layers (Drozdzal et al., 2016; Jégou et al., 2017), our results indicate that information might also be flowing in the other direction i.e., skip and residual connections help layers in the downsampling path to learn spatial information earlier.

Drozdzal et al. (2016) also discuss that layers closer to the center of the model might be more difficult to train due to the vanishing gradient problem and that short skip or residual connections might alleviate this problem. Our results support this as well - middle layers of the SimUnet, which does not have residual or skip connections, seem to learn almost no spatial information compared to the other two networks (**Figure 5A**).

Our results in **Figure 5** also show that models take a largely top-down approach to localizing tumors - they first pay attention to the entire brain, then the general tumor region, and finally converge on the actual finer segmentation. For example, attention in all three models is initially in the background region. In the DenseUnet and ResUnet, attention quickly moves to the brain and whole tumor within the first few layers. Finer segmentations are done in the final few layers. The *necrotic tumor* and *enhancing tumor* are often separated only in the last few layers for all models, indicating that segregating these two regions might require a lesser number of parameters.

This top-down nature is consistent with theories on visual perception in humans–the global-to-local nature of visual perception has been documented. Navon (1977) showed through experiments that larger features take precedence over smaller

features, called the *Global Precedence Effect*. While this effect has its caveats (Beaucousin et al., 2013), it is generally robust (Kimchi, 2015). Brain tumor segmentation models seem to take a similar top-down approach, and we see in our experiments that such behavior becomes more explicit as model performance improves.

While the results from the last two sections are not unexpected, they are not trivial either—the models do not need to learn disentangled concepts, especially implicit ones like the whole brain or the white matter region for which no explicit labels have been given, nor do they need to take a hierarchical approach to this problem. The fact that such human-understandable traces of inference can be extracted from brain tumor segmentation models is promising in terms of their acceptance in the medical domain.

4. EXTRACTING VISUAL REPRESENTATIONS OF INTERNAL CONCEPTS

4.1. Activation Maximization

Visualizing the internal features (i.e., the representations of the internal filters obtained on activation maximization) of a network often provides clues as to the network's understanding of a particular output class. For example, visualizing features of networks trained on the ImageNet (Deng et al., 2009) dataset shows different filters maximally activated either by textures, shapes, objects or a combination of these (Olah et al., 2018). However, this technique has rarely been applied to segmentation models, especially in the medical domain. Extracting such internal features of a brain-tumor segmentation model might provide more information about the qualitative concepts that the network learns and how these concepts develop over layers.

We use the Activation Maximization (Erhan et al., 2009) technique to iteratively find input images that highly activate a particular filter. These images are assumed to be a good first-order representations of the filters. Mathematically, activation maximization can be seen as an optimization problem:

$$x^* = \arg\max_x (\Phi_{k,l}(x) - R_\theta(x) - \lambda||x||_2^2) \tag{6}$$

Where, x^* is the optimized pre-image, $\Phi_{k,l}(x)$ is the activation of the k^{th} filter of the l^{th} layer, and $R_\theta(x)$ are the set of regularizers.

In the case of brain-tumor segmentation, the optimized image is a 4 channel tensor. However, activation maximization often gives images with extreme pixel values or random repeating patterns that highly activate the filter but are not visually meaningful. In order to prevent this, we regularize our optimization to encourage robust images which show shapes and patterns that the network might be detecting.

4.2. Regularization

A number of regularizers have been proposed in the literature to improve the outputs of activation maximization. We use three regularization techniques to give robust human-understandable feature visualizations, apart from an L2 bound which is included in Equation (6).

4.2.1. Jitter

In order to increase translational robustness of our visualizations, we implement Jitter (Mordvintsev et al., 2015). Mathematically, this involves padding the input image and optimizing a different image-sized window on each iteration. In practice, we also rotate the image slightly on each iteration. We find that this greatly helps in reducing high-frequency noise and helps in crisper visualizations.

4.2.2. Total Variation

Total Variation (TV) regularization penalizes variation between adjacent pixels in an image while still maintaining the sharpness of edges (Strong and Chan, 2003). We implement this regularizer to smooth our optimized images while still maintaining the edges. The TV regularizer of an image I with (w, h, c) dimension is mathematically given as in Equation (7):

$$R_{TV}(I) = \sum_{k=0}^{c}\sum_{u=0}^{h}\sum_{v=0}^{w}([I(u,v+1,k) - I(u,v,k)] + [I(u+1,v,k) - I(u,v,k)]) \tag{7}$$

4.2.3. Style Regularizer

In order to obtain visualizations which are similar in style to the set of possible input images, we implement a style regularizer inspired from the work of Li et al. (2017). We encourage our optimization to move closer to the style of the original distribution by adding a similarity loss with a template image, which is the average image taken over the input data distribution. In style transfer, the gram matrix is usually used for this purpose. However, we implement a loss which minimizes the distance between the optimized and template image in a higher dimensional kernel space, as implemented in Li et al. (2017), which is computationally less intensive.

Mathematically, Equation (6) is modified to the following:

$$x^* = \arg\max_x (\Phi_{k,l}(x) - \zeta R_{TV}(x) + \gamma L(x,s) - \lambda||x||_2^2) \tag{8a}$$

$$L(x,s) = \sum_i \sum_j (k(x_i,x_j) + k(s_i,s_j) - 2k(x_i,s_j)) \tag{8b}$$

$$k(x,y) = \exp(-\frac{||x-y||_2^2}{2\sigma^2}) \tag{8c}$$

Where $L(x,s)$ it the style loss between the optimized pre-image and the template image s, $k(x,y)$ is the Gaussian kernel, $\Phi_{k,l}(x)$ is the filter for which activations need to be maximized, $R_{TV}(x)$ is the Total Variation Loss, and $||x||_2^2$ is an upper bound on the optimized pre-image $x*$. Approximate values of the regularization coefficients are $\lambda \sim 10^{-4}$, $\gamma \sim 10^{-2}$, and $\zeta \sim 10^{-5}$. For jitter and rotation, the image is randomly shifted by $\sim$8 pixels, and rotated by $\sim$10 degrees.

The effect of varying the hyperparameters for each of the regularizers is shown in **Supplementary Figure 6**. The effect of jitter is most pronounced—adding jitter by just 2-3 pixels helps reduce high frequency noise and clearly elucidate shapes in the image. Increasing total variation regularization increases smoothness while maintaining shapes and boundaries, reducing

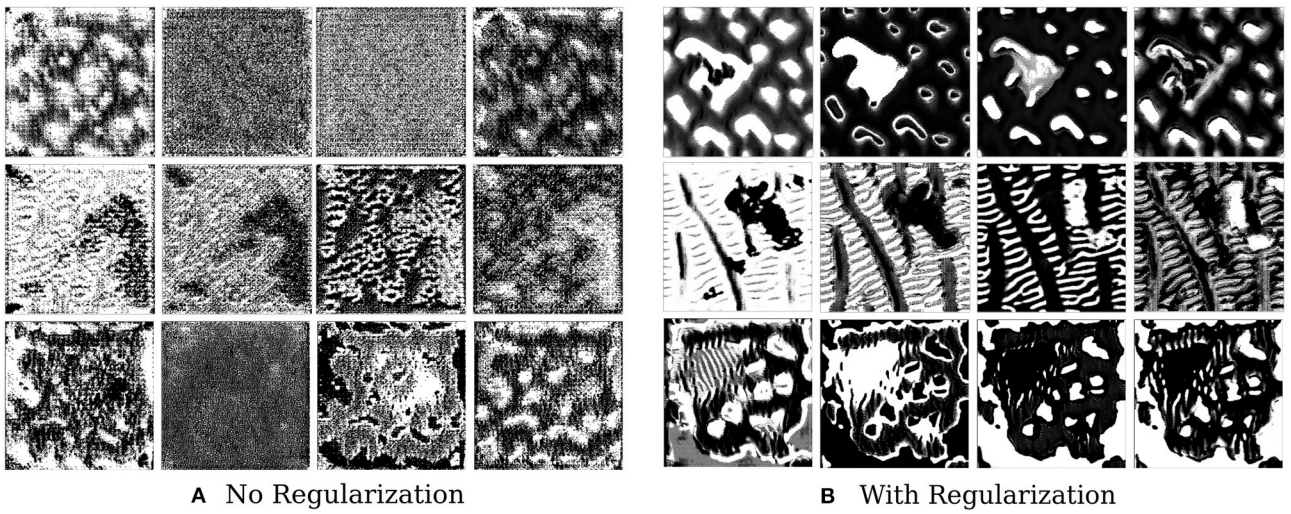


FIGURE 6 | This figure depicts the effect of regularizers on visualized features of brain tumor segmentation models. The four columns on the left show the four channel feature map obtained on optimizing without regularization, while the columns on the right show the effect of adding regularizers.

salt and pepper noise. Increasing style regularization brings the image closer to an elliptical shape similar to a brain. The effect of changing the regularization hyperparameters from a medical perspective in the context brain-tumor segmentation, however, is not clear and further studies would be required in this direction.

We find that style constraining the images and making them more robust to transformations does help in extracting better feature visualizations qualitatively–optimized pre-images do show certain texture patterns and shapes. **Figure 6** shows the results of such an experiment. The effect of regularizers is clear–not regularizing the image leads to random, repeating patterns with high-frequency noise. Constrained images show certain distinct shapes and patterns. It is still not clear, however, that these are faithful reflections of what the filter is actually detecting.

Not a lot of prior work has been done in this area in the context of medical imaging, and our results are useful in the sense that they show that constrained optimization generates such patterns and shapes as compared to noisy unregularized images, which has also been seen in the domain of natural images. In the natural image domain, the resulting pre-images, after regularization, have less high frequency noise and are more easily identifiable by humans. As discussed in the work of Olah et al. (2017) and Nguyen et al. (2016), jitter, L2 regularization, Total Variation, and regularization with mean images priors are shown to produce less noisy and more useful objects or patterns. In medical imaging, however, the resulting patterns and shapes are harder to understand and interpret.

In order to extract clinical meaning from these, a comprehensive evaluation of which regularizers generate medically relevant and useful images based on collaboration with medical professionals and radiologists would be required. This could provide a more complete understanding of what a brain tumor segmentation model actually detects qualitatively. However, this is out of scope of the current study. As we have mentioned in section 7, this will be explored in future work.

5. UNCERTAINTY

Augmenting model predictions with uncertainty estimates are essential in the medical domain since unclear diagnostic cases are aplenty. In such a case, a machine learning model must provide medical professionals with information regarding what it is not sure about, so that more careful attention can be given here. Begoli et al. (2019) discuss the need for uncertainty in machine-assisted medical decision making and the challenges that we might face in this context.

Uncertainty Quantification for deep learning methods in the medical domain has been explored before. Leibig et al. (2017) show that uncertainties estimated using Bayesian dropout were more effective and more efficient for deep learning-based disease detection. Yang et al. (2017) use a Bayesian approach to quantify uncertainties in a deep learning-based image registration task.

However, multiple kinds of uncertainties might exist in deep learning approaches–from data collection to model choice to parameter uncertainty, and not all of them are as useful or can be quantified as easily, as discussed below.

Epistemic uncertainty captures uncertainty in the model parameters, that is, the uncertainty which results from us not being able to identify which kind of model generated the given data distribution. Aleatoric uncertainty, on the other hand, captures noise inherent in the data generating process (Kendall and Gal, 2017). However, Aleatoric Uncertainty is not really useful in the context of this work—we are trying to explain and augment the decisions of the model itself, not the uncertainty in the distribution on which it is fit.

Epistemic uncertainty can, in theory, be determined using Bayesian Neural Networks. However, a more practical and

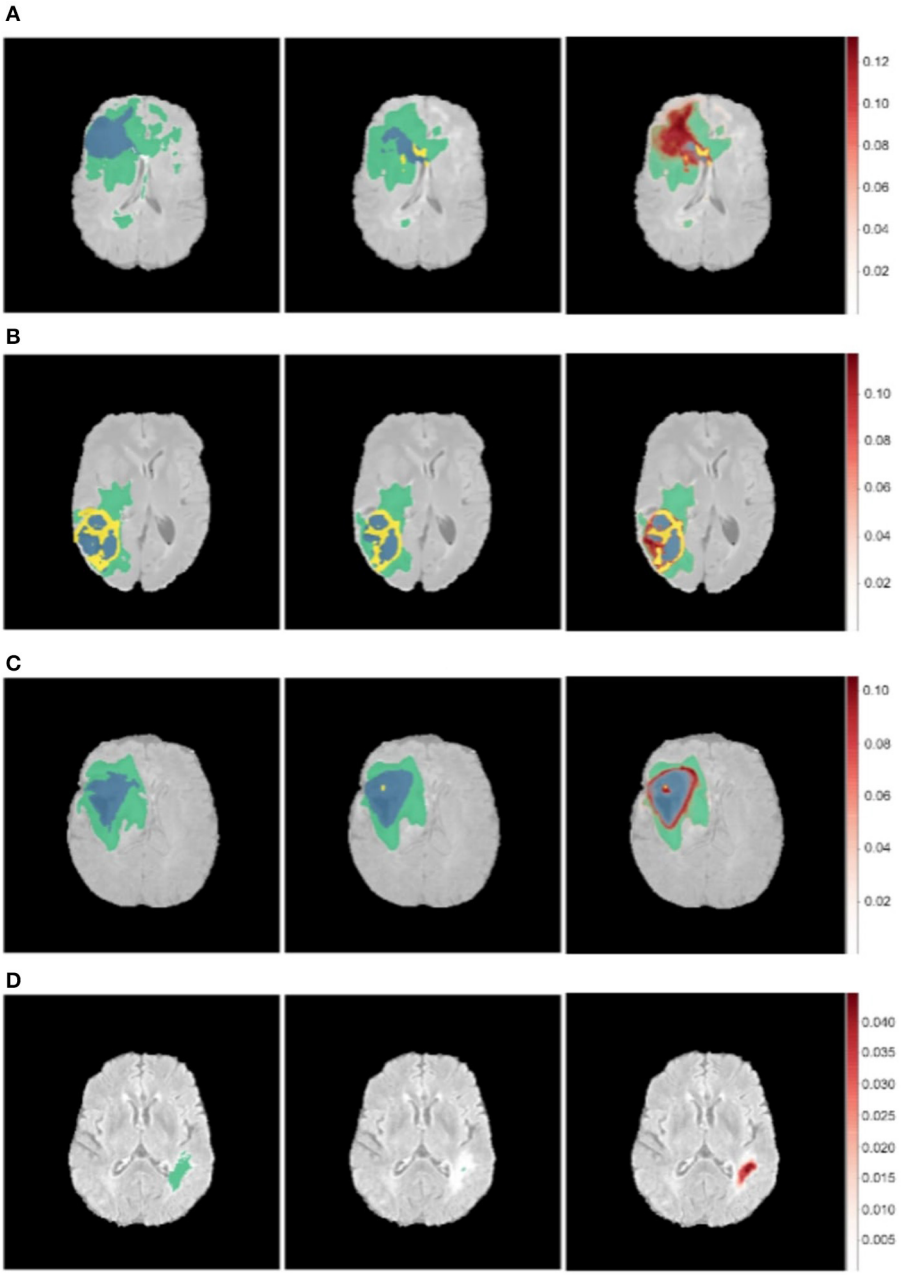


FIGURE 7 | Uncertainty estimations (shown in red) for the DenseUnet using TTD for a selected set of images. Ground Truth (Left), Model Prediction (Middle), and Uncertainty (Right). Misclassified regions are often associated with high uncertainty. **(A)** Misclassified Core Tumor Region which is associated with high model uncertainty. **(B)** Misclassified Enhancing/Core Tumor Region which is associated with high model uncertainty. **(C)** High model uncertainty at class borders. **(D)** Tumor region completely missed by model, captured in the model uncertainty map.

computationally simple approach is to approximate this Bayesian inference by using dropout at test time. We use test time dropout (TTD) as introduced in Gal and Ghahramani (2016) as an approximate variational inference. Then,

$$p(y|x,w) \approx \frac{1}{T}\sum_{t=1}^{t}\Phi(x|w^t) \qquad (9a)$$

$$var_{epistemic}(p(y|x,w)) \approx \frac{1}{T}\sum_{t=1}^{T}\Phi(x|w^t)^T\Phi(x|w^t) - \mathrm{E}(\Phi(x|w^t))^T\mathrm{E}(\Phi(x|w^t)) \qquad (9b)$$

Where $\Phi(x|w^t)$ is the output of the neural network with weights w^t on applying dropout on the t^{th} iteration. The models are retrained with a dropout rate of 0.2 after each layer. At test

time, a posterior distribution is generated by running the model for 100 epochs for each image. We take the mean of the posterior sampled distribution as our prediction and the channel mean of the variance from Equation 9 as the uncertainty (Kendall et al., 2015). The results of this are shown in **Figure 7**.

We find that regions which are misclassified are often associated with high uncertainty. For example, **Figure 7A** shows a region in the upper part of the tumor which is misclassified as *necrotic tumor*, but the model is also highly uncertain about this region. Similar behavior is seen in **Figure 7B**. In some cases, the model misses the tumor region completely, but the uncertainty map still shows that the model has low confidence in this region (**Figure 7D**), while in some cases, boundary regions are misclassified with high uncertainty (**Figure 7C**). In a medical context, these are regions that radiologists should pay more attention to. This would encourage a sort of collaborative effort—tumors are initially segmented by deep learning models and the results are then fine-tuned by human experts who concentrate only on the low-confidence regions, **Figure 1** shows.

More sample images as well as uncertainty for other networks can be found in the **Supplementary Material**.

6. CONCLUSION

In this paper, we attempt to elucidate the process that neural networks take to segment brain tumors. We implement techniques for visual interpretability and concept extraction to make the functional organization of the model clearer and to extract human-understandable traces of inference.
From our introductory study, we make the following inferences:

- Disentangled, human-understandable concepts are learnt by filters of brain tumor segmentation models, across architectures.
- Models take a largely hierarchical approach to tumor localization. In fact, the model with the best test performance shows a clear convergence from larger structures to smaller structures.
- Skip and residual connections may play a role in transferring spatial information to shallower layers.
- Constrained optimization helps to extract feature visualizations which show distinct shapes and patterns which may be representations of tumor structures. Correlating these with the disentangled concepts extracted from Network Dissection experiments might help us understand how exactly a model detects and generalizes such concepts on a filter level.
- Misclassified tumor regions are often associated with high uncertainty, which indicates that an efficient pipeline which combines deep networks and fine-tuning by medical experts can be used to get accurate segmentations.

As we have discussed in the respective sections, each of these inferences might have an impact on our understanding of deep learning models in the context of brain tumor segmentation.

While more experiments on a broader range of models and architectures would be needed to determine if such behavior is consistently seen, the emergence of such human-understandable concepts and processes might aid in the integration of such methods in medical diagnosis–a model which seems to take human-like steps is easier to trust than one that takes completely abstract and incoherent ones. This is also encouraging from a neuroscience perspective - if model behavior is consistent with visual neuroscience research on how the human brain processes information, as some of our results indicate, this could have implications in both machine learning and neuroscience.

7. FUTURE WORK

Future work will be centered around gaining a better understanding of the segmentation process for a greater range of models (including 3D models) and better constrained optimization techniques for extracting human-understandable feature visualizations which would allow an explicit understanding of how models learn generalized concepts. For instance, it would be worth-wile to understand what set of regularizers generates the most medically relevant images. Textural information extracted from the optimized pre-images can also be analyzed to determine their correlation with histopathological features.

Further exploration regarding how these results are relevant from a neuroscience perspective can also be done, which might aid in understanding not just the machine learning model, but also how the brain processes information. The inferences from our explainability pipeline can also be used to integrate medical professionals into the learning process by providing them with information about the internals of the model in a form that they can understand.

DATA AVAILABILITY STATEMENT

Publicly available data sets were used for this study. The data sets can be found at the BRATS 2018 challenge (https://www.med.upenn.edu/sbia/brats2018/data.html) (Bakas et al., 2017a,b).

AUTHOR CONTRIBUTIONS

PN and AK developed the pipeline, performed the analysis, implementation, revised the manuscript, and generated the visualizations. PN wrote the first draft. GK edited the manuscript, supervised, and funded the study.

ACKNOWLEDGMENTS

This work was funded by the Robert Bosch Center for Data Science and Artificial Intelligence (RBCDSAI), under project number CR1920ED617RBCX008562 (Interpretability for Deep Learning Models in Healthcare).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00006/full#supplementary-material

Supplementary Figure 1 | Network Architectures used in our study.

Supplementary Figure 2 | Concepts learned by filters of a particular layer of the ResUnet for an input image (Conv Layer 21).

Supplementary Figure 3 | Concepts learned by filters of a particular layer of the DenseUnet for an input image (Encoding Block 1, Conv 2).

Supplementary Figure 4 | Grad-CAM results for consecutive layers of the ResUnet [view: top to bottom, column **(A)**, followed by top to bottom, column **(B)**].

Supplementary Figure 5 | Activation maps for layers of the ResUnet.

Supplementary Figure 6 | Effect of independently changing hyperparamaters for each regularizer. **(Top)** Jitter coefficient increases [0 pixels, 1p, 6p, 12p, 20p]. **(Middle)** Style Coefficient increases [10^{-2}, 10^{-1}, 1, 5, 10]. **(Bottom)** Total Variation regularization increases [10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}] to smoothen image.

Supplementary Figure 7 | Uncertainty estimations (shown in red) for the DenseUnet **(a–d)** and ResUnet **(e,f)**. Ground Truth (Left), Model Prediction (Middle), and Uncertainty (Right).

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the Tcga-gbm Collection*. The cancer imaging archive (2017).

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). 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.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). "Network dissection: quantifying interpretability of deep visual representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 6541–6549.

Beaucousin, V., Simon, G., Cassotti, M., Pineau, A., Houdé, O., and Poirel, N. (2013). Global interference during early visual processing: Erp evidence from a rapid global/local selective task. *Front. Psychol* 4:539. doi: 10.3389/fpsyg.2013.00539

Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* 1:20. doi: 10.1038/s42256-018-0004-1

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a Large-scale hierarchical image database," in *CVPR09* (IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, eds G. Carneiro, D. Mateus, L. Peter, A. Bradley, J. M. R. S. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, J. S. Cardoso, and J. Cornebise (Athens: Springer), 179–187. doi: 10.1007/978-3-319-46976-8_19

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *Univ. Montreal* 1341:1. Available online at: https://www.semanticscholar.org/paper/Visualizing-Higher-Layer-Features-of-a-Deep-Network-Erhan-Bengio/65d994fb778a8d9e0f632659fb33a082949a50d3#paper-header

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning* (New York, NY), 1050–1059.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). "Explaining explanations: an overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (Turin: IEEE), 80–89.

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv preprint* arXiv:1712.09923.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). "The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Honolulu, HI), 11–19.

Kanwisher, N., and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. B Biol. Sci.* 361, 2109–2128. doi: 10.1098/rstb.2006.1934

Kendall, A., Badrinarayanan, V., and Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint* arXiv:1511.02680.

Kendall, A., and Gal, Y. (2017). "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA), 5574–5584.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 1122–1131. doi: 10.1016/j.cell.2018.02.010

Kermi, A., Mahmoudi, I., and Khadir, M. T. (2018). "Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Granada: Springer), 37–48. doi: 10.1007/978-3-030-11726-9_4

Kimchi, R. (2015). *The Perception of Hierarchical Structure*. Oxford handbook of perceptual organization, 129–149.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.* 7:17816. doi: 10.1038/s41598-017-17876-z

Li, Y., Wang, N., Liu, J., and Hou, X. (2017). Demystifying neural style transfer. *arXiv preprint* arXiv:1701.01036.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imag.* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Molnar, C. (2018). *Interpretable Machine Learning*. A Guide for Making Black Box Models Explainable, 7.

Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: Going deeper into neural networks. *Google AI Blog* (Google). Available online at: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cogn. Psychol.* 9, 353–383.

Nguyen, A., Yosinski, J., and Clune, J. (2016). Multifaceted feature visualization: uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint* arXiv:1602.03616.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*. Available online at: https://distill.pub/2017/feature-visualization (accessed August 30, 2019).

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., et al. (2018). The building blocks of interpretability. *Distill*. Available online at: https://distill.pub/2018/building-blocks (accessed August 28, 2019).

Pitcher, D., Charles, L., Devlin, J. T., Walsh, V., and Duchaine, B. (2009). Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Curr. Biol.* 19, 319–324. doi: 10.1016/j.cub.2009.01.007

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 618–626.

Shaikh, M., Anand, G., Acharya, G., Amrutkar, A., Alex, V., and Krishnamurthi, G. (2017). "Brain tumor segmentation using dense fully convolutional neural network," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Quebec City, QC: Springer), 309–319. doi: 10.1007/978-3-319-75238-9_27

Strong, D., and Chan, T. (2003). Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Problems* 19:S165. doi: 10.1088/0266-5611/19/6/059

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 12:e0174944. doi: 10.1371/journal.pone.0174944

Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Quicksilver: fast predictive image registration–a deep learning approach. *NeuroImage* 158, 378–396. doi: 10.1016/j.neuroimage.2017.07.008

Zamir, A. R., Wu, T.-L., Sun, L., Shen, W. B., Shi, B. E., Malik, J., et al. (2017). "Feedback networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (San Juan, Puerto Rico), 1308–1317.

Zeki, S., and Bartels, A. (1998). The autonomy of the visual systems and the modularity of conscious vision. *Philos. Transact. R. Soc. Lond. Ser. B Biol. Sci.* 353, 1911–1914. doi: 10.1098/rstb.1998.0343

Zhang, Q.-S., and Zhu, S.-C. (2018). Visual interpretability for deep learning: a survey. *Front. Inform. Techn. Electr. Eng.* 19, 27–39. doi: 10.1631/FITEE.1700808

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2014). Object detectors emerge in deep scene cnns. *arXiv preprint* arXiv:1412.6856.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 07 February 2020
doi: 10.3389/fnins.2020.00065

Systematic Evaluation of Image Tiling Adverse Effects on Deep Learning Semantic Segmentation

G. Anthony Reina[1], *Ravi Panchumarthy*[1], *Siddhesh Pravin Thakur*[2], *Alexei Bastidas*[1] *and Spyridon Bakas*[2,3,4*†]

[1] *Intel Corporation, Santa Clara, CA, United States,* [2] *Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA, United States,* [3] *Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States,* [4] *Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States*

Convolutional neural network (CNN) models obtain state of the art performance on image classification, localization, and segmentation tasks. Limitations in computer hardware, most notably memory size in deep learning accelerator cards, prevent relatively large images, such as those from medical and satellite imaging, from being processed as a whole in their original resolution. A fully convolutional topology, such as U-Net, is typically trained on down-sampled images and inferred on images of their original size and resolution, by simply dividing the larger image into smaller (typically overlapping) tiles, making predictions on these tiles, and stitching them back together as the prediction for the whole image. In this study, we show that this tiling technique combined with translationally-invariant nature of CNNs causes small, but relevant differences during inference that can be detrimental in the performance of the model. Here we quantify these variations in both medical (i.e., BraTS) and non-medical (i.e., satellite) images and show that training a 2D U-Net model on the whole image substantially improves the overall model performance. Finally, we compare 2D and 3D semantic segmentation models to show that providing CNN models with a wider context of the image in all three dimensions leads to more accurate and consistent predictions. Our results suggest that tiling the input to CNN models—while perhaps necessary to overcome the memory limitations in computer hardware—may lead to undesirable and unpredictable errors in the model's output that can only be adequately mitigated by increasing the input of the model to the largest possible tile size.

Keywords: segmentation, tiling, deep learning, CNN, brain tumor, glioma, BraTS, satellite imaging

OPEN ACCESS

Edited by:
John Ashburner,
University College London,
United Kingdom

Reviewed by:
Hannah Spitzer,
Julich Research Centre, Germany
Guodong Zeng,
University of Bern, Switzerland

***Correspondence:**
Spyridon Bakas
sbakas@upenn.edu

†ORCID:
Spyridon Bakas
orcid.org/0000-0001-8734-6482

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *29 August 2019*
Accepted: *16 January 2020*
Published: *07 February 2020*

Citation:
Reina GA, Panchumarthy R, Thakur SP, Bastidas A and Bakas S (2020) Systematic Evaluation of Image Tiling Adverse Effects on Deep Learning Semantic Segmentation. Front. Neurosci. 14:65. doi: 10.3389/fnins.2020.00065

1. INTRODUCTION

Since their resurgence in 2012 convolutional neural networks (CNN) have rapidly proved to be the state-of-the-art method for computer-aided diagnosis in medical imaging, and have led to improved accuracy in classification, localization, and segmentation tasks (Krizhevsky et al., 2012; Chen et al., 2016; Greenspan et al., 2016). However, memory constraints in deep learning accelerator cards have often limited training on large 2D and 3D images due to the size of the

activation maps held for the backward pass during gradient descent (Chen et al., 2016; Ito et al., 2019). Two methods are commonly used to manage these memory limitations: (i) images are often down-sampled to a lower resolution, and/or (ii) images are broken into smaller tiles (Huang et al., 2018; Pinckaers and Litjens, 2018). Tiling is often applied when using large images due to the memory limitations of the hardware (Roth et al., 2018). Specifically, in CNN models, the activation maps of the intermediate layers use several times the memory footprint of the original input image. These activation maps can easily increase the allocated memory to hundreds of gigabytes. Fully convolutional networks are a natural fit for tiling methods, as they can be trained on images of one size and perform inference on images of a larger size by breaking the large image into smaller, overlapping tiles (Ronneberger et al., 2015; Çiçek et al., 2016; Roth et al., 2018). To perform the overlapping tiling at inference time, varying $N \times N$ (or in the 3D case, $N \times N \times N$) tiles are cropped from the whole image at uniformly spaced offsets along the image dimensions.

Tiling introduces additional model hyperparameters—namely, tile size, overlap amount, and aggregation process (e.g., tile averaging/rounding)—that must be tuned to generate better predictions. For example, Roth et al. performed abdominal organ segmentation on 512×512 CT images with between 460 and 1,177 slices by using input tiles of size $132 \times 132 \times 116$ to yield output prediction tiles of $44 \times 44 \times 28$ in a Cascaded 3D U-Net (Roth et al., 2018). In the second stage of the prediction, the probabilities for overlapping tile predictions were averaged to produce a better *Dice* Coefficient result. Zeng and Zheng (2018) introduced "Holistic Decomposition Convolution" that—when added to a conventional 3D U-Net—significantly reduced the size of the input data while maintaining the useful information for the semantic segmentation. They compared the effects of $50 \times 50 \times 40$, $96 \times 96 \times 96$, and $200 \times 200 \times 40$ tile crops from a $480 \times 480 \times 160$ MR and determined that they had better *Dice* Coefficient, Hausdorff Distance, and Average Surface Distance when using the largest tile size that could fit into memory. Isensee et al. (2019) used a sliding window with a half-tile overlap and test-time data augmentation that mirrored the tile along all axes. They also favored larger tile size over large batch size in order to "maximize the amount of spatial context that can be captured." Ghosh et al. (2018) found that by rotating or flipping the input tile, the prediction was slightly different for the same tile. By averaging these small variations in the tiled predictions, Ghosh produced improved predictions in structures within satellite imagery from a dilated U-Net topology. Huang et al. determined that zero-padding and strided convolutions (i.e., stride > 1)—two methods commonly used in CNNs—created variability in predictions close to the tile border and caused translation variance in the output prediction (Huang et al., 2018).

Previous works like these refer to tiling methods as "necessary due to constraints in memory" rather than methods to "improve the accuracy of the algorithms" (Chen et al., 2016; Roth et al., 2018; Isensee et al., 2019; Ito et al., 2019). In other words, the tiling method compensates for insufficient memory rather than adds predictive power. If more memory were available for training and inference of these models, then tiling methods would have not been necessary or even desirable. For example, Kamnitsas et al. (2017) created the first state of the art 3D topology for predicting brain tumors by finding tiles of "image-segments" which are "larger than individual patches [tiles], but small enough to fit into memory." Roth et al. (2018) remarked, "with the growing amount of …memory, overlapping sub-volume predictions …will be reduced as it will be come possible to reshape the network to accept arbitrary 3D input image sizes."

In this study, we focus on the tiling approach—during both model training and model inference—and its influence on the model prediction. We implemented U-Net topologies for both 2D (Ronneberger et al., 2015) and 3D (Çiçek et al., 2016) data, and we question whether this image tiling approach is indeed as accurate as simply performing inference on the whole image. In a previous report (Reina and Panchumarthy, 2018), we noticed that using the entire 2D image gave better predictions than the tiling approach for a 2D U-Net model trained to detect glial tumors from brain magnetic resonance imaging (MRI). In this study, we extend those results by systematically (i) evaluating the resulting effects in both medical and non-medical data, (ii) comparing both 2D and 3D U-Net models, and (iii) suggesting that these differences are caused by operations within the CNN model that vary due to translations in the input of the model. Finally, we show that these issues can be partially addressed by increasing the size of the tile—up to and including training and inferring on the whole image.

2. METHODS

2.1. Data

2.1.1. Brain Tumor Segmentation (BraTS)

The medical data used for our evaluations reflect the publicly-available training dataset of the International Brain Tumor Segmentation (BraTS) challenge 2019[1] (**Figure 1**) (Menze et al., 2014; Bakas et al., 2017a,b,c; Bakas et al., 2018). BraTS created a publicly-available multi-institutional dataset for benchmarking and quantitatively evaluating the performance of computer-aided segmentation algorithms for brain tumors from MRI scans. These scans were acquired by 1T, 1.5T, or 3T MRI scanners and all the ground truth labels were manually approved by expert, board-certified neuroradiologists. The dataset we used here comprises pre-operative multi-parametric MRI scans from 335 patients diagnosed with glioma. The exact modalities of the mpMRI scans included describe native T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR) scans. We randomly split this dataset into 270 training, 30 validation, and 35 testing scans.

Although the BraTS data describe 3D MRI scans, here we are considering the 155 2D slices from each scan to be an independent image for training a 2D model. However, all 2D slices from a single patient scan were contained in only one of the three dataset splits (training/validation/testing), to prevent any potential data leakage toward learning data co-linearities.

[1] www.med.upenn.edu/cbica/brats2019.html

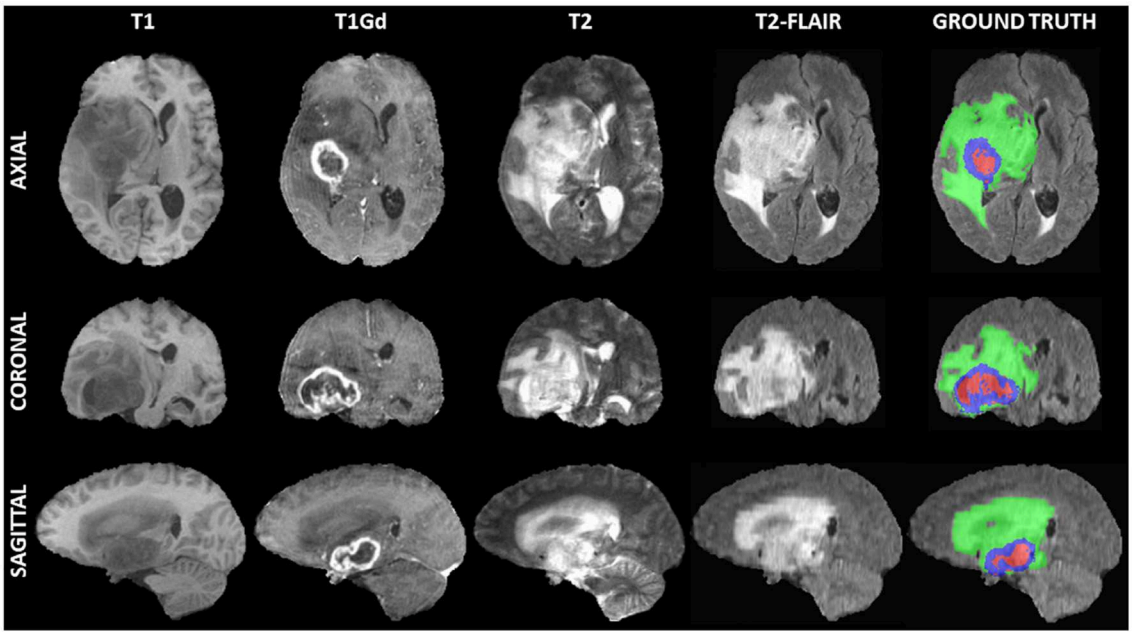


FIGURE 1 | Example of a 3D input multi-parametric Magnetic Resonance Imaging scan from the International Brain Tumor Segmentation (BraTS) challenge. From left to right all four input modalities are illustrated, including native T1-weighted (T1), T1 post-contrast (T1Gd), native T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR), followed by the ground truth expert annotation of all three tumor sub-regions, provided as part of the BraTS dataset. From top to bottom three views (i.e., Axial, Coronal, Sagittal) of these 3D volumes are depicted to showcase the 3-dimensional nature of these scans.

Specifically, there were 41,850, 4,650, and 5,425 2D image/mask pairs corresponding to 270, 30, and 35 3D MRI scans, across the training, validation, and testing sets, respectively. All 2D images were Z-scored along the channel axis from pre-computed means and standard deviations of the 3D MRI scan. The original 2D slices were 240 × 240 pixels (i.e., whole image).

2.1.2. SpaceNet Vegas Satellite Imagery

The non-medical data is sourced from the public SpaceNet satellite imagery dataset suite (**Figure 2**) (SPA, 2018; Weir et al., 2019)[2]. Specifically, we used the Vegas subset of the data (SN-Vegas). It is comprised of 3,851 30 cm spatial resolution, pan-sharpened, RGB satellite imagery over the city of Las Vegas, Nevada (USA) as well as latitude-longitude annotations for 108,942 building footprint polygons within the city. We exclude the official competition test dataset from this study because it does not contain publicly-available ground truth annotations. The images were captured by WorldView-2 and 3 satellites, and filtered to exclude images with excessive cloud cover as well as extreme capture angles. The labels were professionally created by geospatial data labeling vendor Radiant Solutions[3].

The SpaceNet-Vegas dataset was split into 70% training (2,695 images), 20% validation (770 images), and 10% testing (386 images), corresponding to 77,099 training, 21,505 validation, and 10,338 testing building polygons. All inputs were Z-scored along the channel axis from pre-computed means and standard deviations. All training inputs were also subject to random horizontal and vertical flips, and rotations between 0 and 360°.

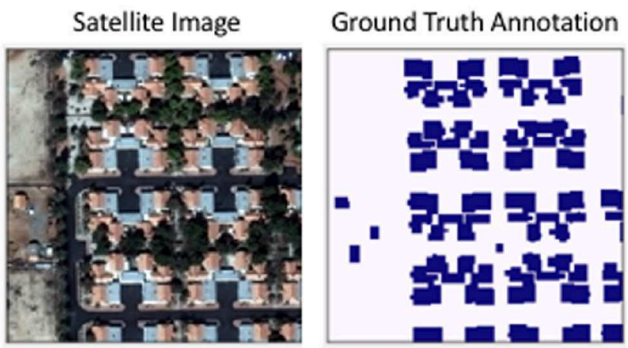


FIGURE 2 | An example of the SpaceNet-Vegas images used in this study. The ground truth annotations for buildings and other structures were professionally labeled.

2.2. U-Net Topology

U-Net is a fully convolutional network based on an encoder-decoder architecture (**Figure 3**). The contracting path captures context and the expanding path enables localization. Unlike the standard encoder-decoder, each feature map in the expanding path is concatenated with a corresponding feature map from the contracting path, augmenting downstream feature maps

[2]spacenet.ai/spacenet-buildings-dataset-v2/
[3]www.radiantsolutions.com

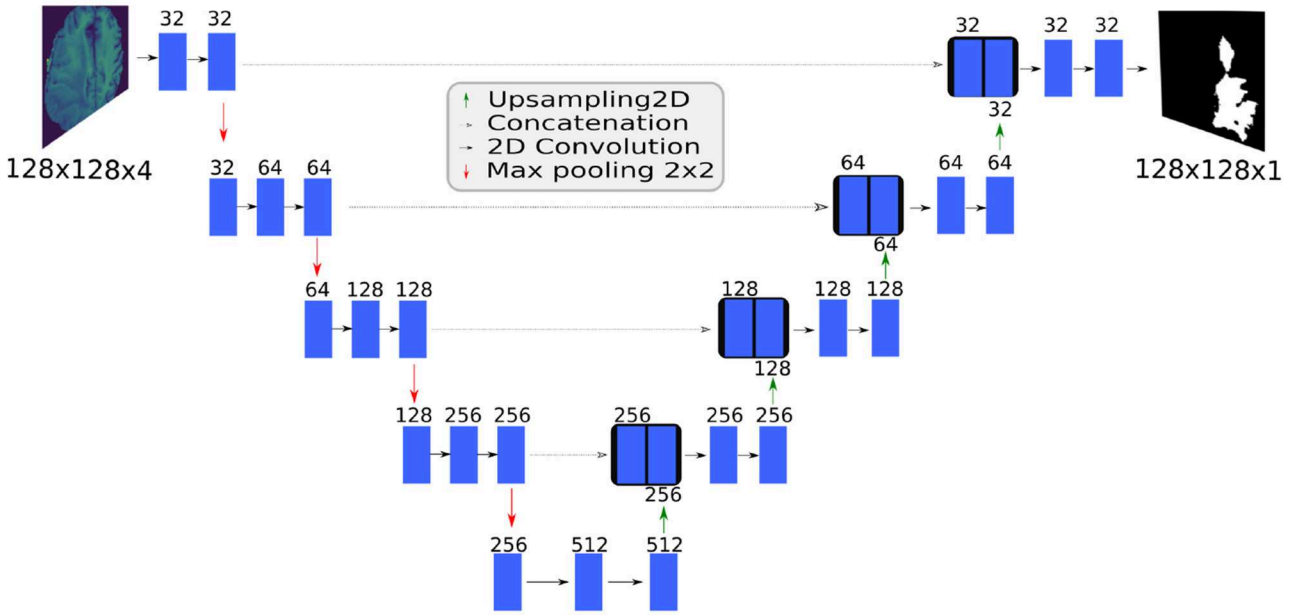


FIGURE 3 | The BraTS 2D U-Net topology. The SpaceNet 2D U-Net and the BraTS 3D U-Net topologies have similar architectures.

with spatial information acquired using smaller receptive fields. Intuitively, this allows the network to consider features at various spatial scales. By design, U-Net is agnostic to image size, and its training and inference can be performed on images of different size.

2.3. 2D U-Net for Medical Data (BraTS)

We adapted a 2D U-Net model for training on the BraTS data, and specifically used four MRI modalities as input and output an equivalently-sized mask predicting the whole tumor appearing in a 2D slice.

2.3.1. Architectural Modifications

In favor of allowing wider reproducibility of our results, we specifically modified the originally published 2D U-Net topology by reducing the number of feature maps by half (from 64 in the first convolutional layer down to 32) and adding dropout (0.2) just before the 3rd and 4th max pooling layers. We also used zero padding in all convolutional layers to maintain the image dimensions and eliminate the need to crop the image for concatenation. The reduction of the originally proposed feature maps happened in favor of our results been reproducible by others without requiring extreme hardware equipment.

2.3.2. Training Process

We implemented the model used here in Keras 2.2.4 and TensorFlow 1.11, and made the complete source code publicly available[4]. Stochastic gradient descent with the Adam optimizer (learning rate = 1e-4) was used to minimize the loss function $-\log(Dice)$, where *Dice* is defined as in equation 1 on page 6. A batch size of 128 was used during training. We created a batch generator which randomly selected cropped images/masks from the training set for each batch.

The 2D model was trained for 40 epochs. During training, a random crop of 128 × 128 pixels was taken from the normalized 2D images and their corresponding ground truth masks. Randomized flipping (up/down and left/right), and 90 degree rotation of the training set images were also used during online data augmentation. The *Dice* on a center 128 × 128 crop of the validation dataset was calculated after every epoch. The model that produced the highest *Dice* on the center 128 × 128 crop of the validation data was considered the best trained model.

For pre-processing of the images, on a per image basis, images were clipped to 98 percentile of their values and standardization was applied only on non-zero pixels making background consistent over all images. This created a consistent effect of normalization over the images.

2.3.3. Zero Padding Experiments

We conducted additional experiments to determine the effects of zero padding on the tiling approach. This stemmed from the findings of Huang et al. (2018), who suggested that zero padding used in CNN topologies caused variability in predictions at the tile border. To assess this, we also created and trained an additional 2D U-Net model that did not include zero-padding for any of the convolutional layers. We named this the "no pad BraTS" model.

The "no pad BraTS" model was trained in the same way as the first 2D U-Net model, but with the following changes. This "no pad BraTS" model took as input a random crop of 236 × 236 and output a 52 × 52 prediction. The decrease in the output size was due to the progressive loss in the border pixels after

[4]github.com/IntelAI/unet

each non-padded convolutional layer (Ronneberger et al., 2015). The input size was chosen to be slightly smaller than the whole 240×240 slice so that we could evaluate if the prediction changed with small translations of the input. The model prediction was compared to a similarly-cropped version of the ground truth mask. It was trained for 40 epochs and the model that produced the highest *Dice* on the validation data was considered the best trained model.

2.3.4. Inferring on 2D Tiles (Tiling Approach)

Inference was performed individually on five 128×128 pixel-sized tiles, extracted from the four corners and the center of the slice (**Figure 4A**). We performed inference on the whole 2D slice using the model and then stacked the 155 slices on a per scan basis to generate a predicted 3D segmentation mask of the entire scan.

We utilized and compared two tiling aggregation approaches. The first approach, *rounding after averaging*, is described by Roth et al. (2018). In our case the predictions from these five 128×128 tiles were first averaged and then rounded to either 0 or 1 (threshold:0.5). We compared the *rounding after averaging* approach with a *rounding before averaging* approach: The five 128×128 tiles were rounded to either 0 or 1 (threshold:0.5) and then averaged to provide the whole image prediction (*rounding before averaging*). Slicewise predictions for each patient scan were then stacked together to compare the 2D predictions with predictions from the 3D BraTS model.

2.3.5. Inferring on the Whole 2D Slice

For fully-convolutional topologies, the TensorFlow model can be created with a run-time defined height and width by specifying the input dimensions to be $[Height, Width, Channels] = [None, None, 4]$ where *None* describes the run-time defined parameter[5].

By defining and training the model in this manner, we can pass an image of almost any size into the model and perform inference. The only limitation to the input image size is that the dimension must be divisible by 2^4 in order to align with the 4 max-pool layers of the U-Net model and correctly concatenate the skip connections.

2.4. 3D U-Net for Medical Data (BraTS)

To create the 3D U-Net model, we used the same number of convolutional and max-pooling layers as we used in the 2D U-Net model (**Figure 3**). We altered the implementation of the originally proposed 3D U-Net model (Çiçek et al., 2016) by replacing the ReLU layer with a leaky ReLU activation and adding instance normalization after each leaky ReLU (Xu et al., 2015; Ulyanov et al., 2016).

We further modified this implementation by using an initial learning rate of 0.01. A learning rate decay factor of 0.5 was applied when the value of the validation loss had not been in the five best previous losses (i.e., check_best = 5). Training stopped when the validation loss did not improve in the past 20 epochs (i.e., patience = 20). Finally, the weights that yielded the lowest validation loss were used for the final model.

[5]inputs = tensorflow.keras.layers.Input([None, None, number_channels_in]).

The 3D BraTS model is trained for 100 epochs on 9 tiles, of $128 \times 128 \times 128$ voxels, cropped from the 8 corners and the center of a 3D MRI scan (**Figure 4B**).

Inferring via a tiling approach was also performed similar to the 2D U-Net case (section 2.3.4), but used $128 \times 128 \times 128$ tiles from the eight corners and the center of the whole image (**Figure 4B**). These nine $128 \times 128 \times 128$ tiles were averaged to provide a prediction of the whole mask.

Whole image inference was also performed similar to the 2D U-Net (section 2.3.5) but using the whole $240 \times 240 \times 155$ scan.

2.5. 2D U-Net on Satellite Data (SpaceNet-Vegas)

The SpaceNet model uses a single satellite image from SpaceNet-Vegas as input, and outputs an equivalently-sized mask predicting the building footprints.

2.5.1. Architectural Modifications

The originally published topology was modified by introducing batch normalization to the output of a convolution layer, prior to the activation, for regularization purposes.

2.5.2. Training Process

All models were trained for 300 epochs using the Adam optimizer with a 5e-4 learning rate to optimize the Binary Cross Entropy loss. To test our hypothesis that a model trained in the whole image outperforms a tiling-based approach, we followed two training processes here; based on (a) tiling, and (b) down-sampling.

For models trained via tiling, the input image's source resolution of 650×650 is maintained and a random crop of the desired dimension is selected. Different models were trained for each of the following random tiling sizes:

- 128×128
- 256×256
- 384×384
- 496×496

Due to the U-Net architecture, the input dimensions to the model must be divisible by 2^5 in order to align with the 5 max-pool layers. Consequently, for the models trained on the entire image via down-sampling, the original image was downsampled with anti-aliasing and bilinear interpolation to 512×512 and 640×640.

2.5.3. Zero Padding Experiments

As with the BraTS experiments, we created additional SpaceNet experiments to determine the effects that zero padding had on the tiling approach. We also created and trained additional SpaceNet models that did not include zero-padding for any of the convolutional layers (namely the "no pad SpaceNet" models).

2.5.4. Inferring on 2D Tiles

Inference was performed using tiles of the same size that was used when training the model, with a 50% overlap between tiles in both the vertical and horizontal dimension. The overlapping tiles were averaged to provide the whole image prediction.

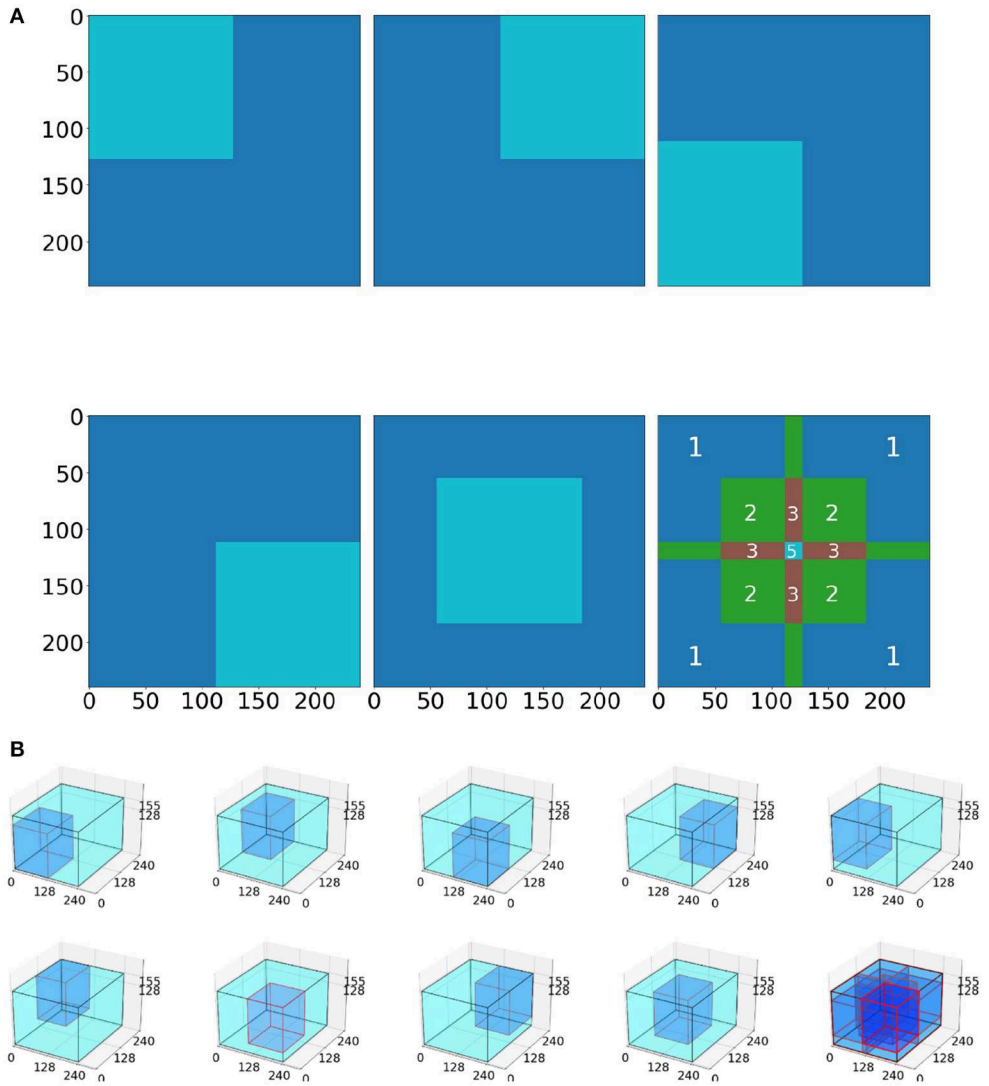


FIGURE 4 | Tiling process schematic. **(A)** In the 2D model, five tiles (4 corners, 1 center) are averaged to produce the whole image prediction. The 3rd picture in the 2nd row depicts the intensity of the tile overlapping. Notably, the tile predictions are either (i) first rounded and then averaged together, or (ii) first averaged together and then rounded. **(B)** Example for the 3D BraTS model, where the tiling algorithm is similar with the 2D, but this time uses nine tiles (8 corners and 1 center).

2.6. Evaluation Metric

2.6.1. …for the Medical Data

In consistency with the metric used in the BraTS challenge, the Dice Similarity Coefficient (*Dice*) was used here to measure the quality of the tumor predictions. *Dice* is defined as:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (1)$$

where TP, FP, TN, FN are the number of True Positive, False Positive, True Negative, and False Negative pixels.

2.6.2. …for the Satellite Data

In order to measure performance relative to established benchmarks on SpaceNet, we used the post-processing Polygon *F*1 metric, displayed in **Figure 5**; namely, the predicted segmentation mask is polygonized based on same-value pixel connectivity to generate a set of proposed polygons in latitude and longitude space. We then calculate the spatial intersection over union (i.e., Jaccard Index) between proposed and ground truth polygons. A true positive is asserted if the Jaccard value is above 0.5. Once we establish TP, FP, and FN counts, we compute the *Dice* (also known as SpaceNet (polygonal) *F*1 Score)—the harmonic mean between precision and recall—over these matched polygons and compare this metric to the *Dice* calculated on a pixelwise basis (Hagerty, 2016).

3. RESULTS

3.1. 2D BraTS Model

The best trained 2D BraTS model yielded an average *Dice* of 0.8877 when inferred on a single center 128 × 128 tile of the test dataset slices. Furthermore, as explained in the methods,

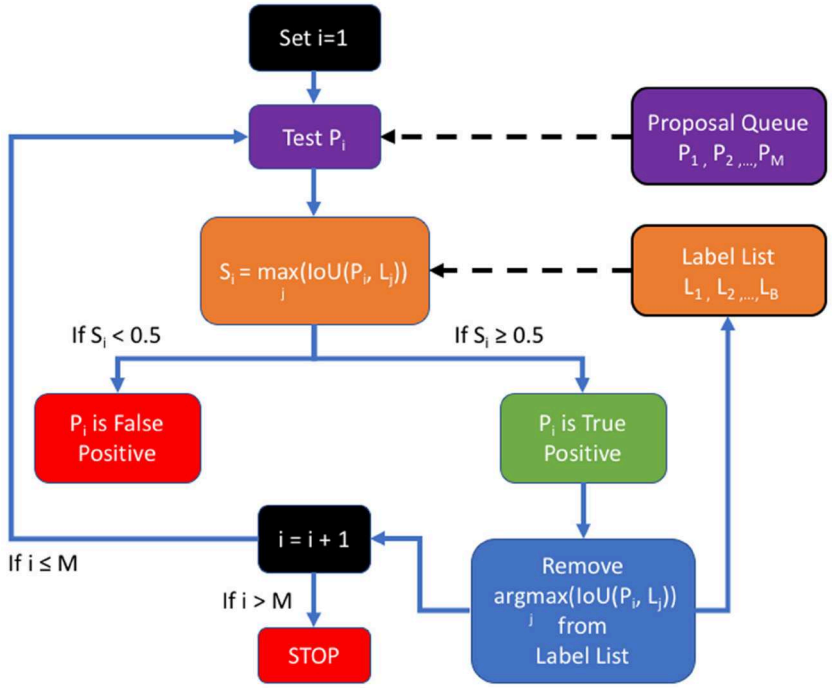


FIGURE 5 | The SpaceNet *F*1 metric: a list of proposals is generated by the detection algorithm and compared to the ground truth in the list of labels.

although this model was trained on random 128 × 128 tiles, we were able to perform inference on the entire 240 × 240 2D image slice. The whole 2D slice predictions resulted in an average *Dice* on 0.8743 on the whole 3D volume. Using the 2D BraTS model with five 128 × 128 tiles, resulted in an average *Dice* of 0.8599 for the tiling aggregation method of *rounding after averaging*. Using the *rounding before averaging* tiling aggregation method, resulted in a 0.8998 average *Dice* (**Table 1**).

Collectively in the testing dataset, application of different tiling aggregation approaches (i.e., *rounding after averaging*, and *rounding before averaging*) revealed that when we aggregated the predicted segmentations by *rounding after averaging*, the 2D segmentations of individual subjects were inferior to the segmentations obtained from the 3D model. Contrarily, evaluation of the tiling aggregation approach, where *rounding before averaging* was applied, yield that on average more 2D predictions were closer to the ground truth than when using the 3D model (**Figure 6**).

3.2. 3D BraTS Model

The results of the 3D U-Net BraTS model showed a different behavior when compared with the results of the 2D U-Net model. Specifically, there were no significant differences observed when the predictions of the model inferred on the whole 3D MRI scan were compared to the predictions of any of the tiling aggregation approaches. Inferring the 3D BraTS model on the whole 3D scan resulted in an average *Dice* of 0.8974, when for the tiling aggregation method of *rounding after averaging* and of *rounding before averaging* the average *Dice* was equal to 0.8991 and 0.8984, respectively (**Table 2**).

TABLE 1 | Results of 2D U-Net on medical data (BraTS).

Inference on:	Whole 2D slice	2D tiles	2D tiles
Aggregation approach	N/A	(Rounding after averaging)	(Rounding before averaging)
Dice	0.8743	0.8599	0.8998

Comparing whole 2D slice prediction to two tiling aggregation methods.

3.3. 2D SpaceNet-Vegas

With the satellite image dataset, we note that higher accuracy was obtained by training on a larger tile size (i.e., larger context of the image). The model trained on 128 × 128 random tiles and inferred on the whole 650 × 650 image with 128 × 128 sliding tiles, resulted in a *Dice* score of 0.791, whereas the model trained on the whole 2D image resized to 640 × 640 and inferred on the whole 650 × 650 image resulted in a *Dice* score of 0.917. To train on the whole image, we interpolated the image to 640×640 as the U-Net topology require the input image to be multiple of 2^5 to align with 5 max-pool layers. **Tables 3**, **4** denote that both the evaluation metrics of *Dice* and SpaceNet *F*1 (polygon-wise and computed over the entire dataset, not per image) improve as the training tile size increases.

4. DISCUSSION

Our results denote substantial differences in our 2D U-Net architecture, both for medical and non-medical (i.e., satellite) data. Specifically, the evaluation of *Dice* show superiority when

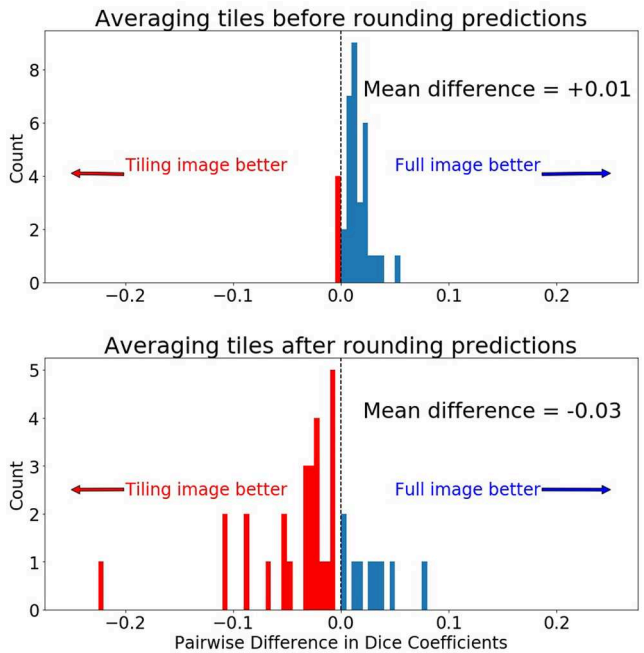


FIGURE 6 | Comparing pairwise *Dice* differences between the prediction on the whole 240 × 240 image and the 128 × 128 tiles. Zero indicates both methods produced equal *Dice* scores for the same scan. **(Top)** Prediction was made by first averaging the 5 tiles and then rounding the final prediction to 0 or 1. **(Bottom)** Prediction was made by first rounding the tiled predictions to 0 or 1 and then averaging the predictions.

TABLE 2 | Results of 3D U-Net on medical data (BraTS).

Inference on:	Whole 3D scan	3D tiles	3D tiles
Aggregation approach	N/A	(Rounding after averaging)	(Rounding before averaging)
Average *Dice* ($\pm\sigma$)	0.8974 ($\pm$ 0.0702)	0.8991 ($\pm$ 0.0666)	0.8984 ($\pm$ 0.0670)

Comparing whole 3D scan prediction to two tiling aggregation methods.

TABLE 3 | Results of 2D U-Net with zero-padding on non-medical data (SpaceNet Vegas).

Tile size	128 × 128 crop	256 × 256 crop	384 × 384 crop	496 × 496 crop	512 × 512 interp	640 × 640 interp
Dice	0.873	0.900	0.896	0.918	0.917	0.918
SpaceNet *F*1	0.748	0.803	0.800	0.838	0.840	0.847

Dice and SpaceNet (polygon-wise Dice) F1 metrics on varying tile size.

inferring our model in the whole 2D image, when compared with inferring in smaller image tiles, supporting our hypothesis for the large tile sizes. Furthermore, gradual increments of the tile sizes shows gradual improvement in the performance. Following the evaluation of our 3D U-Net model, we note that the performance on 3-dimensional data did not show substantial difference when comparing inference on the whole 3D image and inference on

TABLE 4 | Results of 2D U-Net without zero-padding on non-medical data (SpaceNet Vegas).

Tile size	128 × 128 crop	256 × 256 crop	384 × 384 crop	496 × 496 crop	512 × 512 interp	640 × 640 interp
Dice	0.865	0.896	0.907	0.918	0.912	0.914
SpaceNet *F*1	0.734	0.781	0.797	0.806	0.808	0.821

Dice and SpaceNet (polygon-wise Dice) F1 metrics on varying tile size.

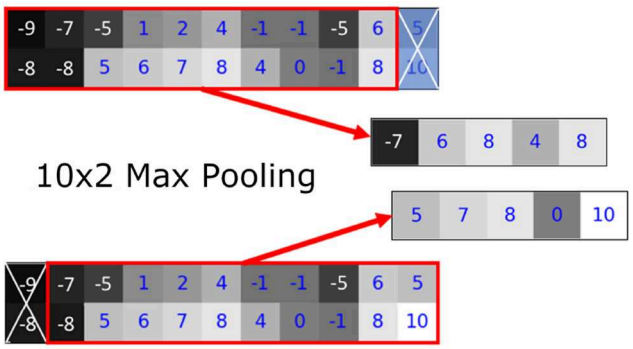


FIGURE 7 | The translationally-variant nature of the MaxPooling layer: note how the result of MaxPooling significantly differs with a translation of one pixel to the 10 × 2 window between the top and bottom inputs even though they contain the same values (middle rows). Successive MaxPooling layers (or any non-unary, strided convolution) compound the effect because they effectively increase the receptive field window size. Hence, a model with three 2 × 2 MaxPooling layers would show translational variance for offsets of up to $2^3 = 8$ pixels.

3D tiles. We hypothesize that this happens due to the inclusion of large image context (e.g., more neighboring voxels) along the third dimension.

The overlapping tiling approach is commonly used by researchers to apply fully convolutional models on large 2D and 3D images that would ordinarily not fit into available memory (Chen et al., 2016; Roth et al., 2018). Isensee et al. (2019), for example, specifically designed their topology to "automatically set the batch size, tile size and number of pooling operations for each axis while keeping the memory consumption within a certain budget." We suggest that researchers should be designing their topologies not to fit into a hardware constraint, but instead to produce the most accurate model possible.

We found that the variance in the prediction can be seen in the linear transformation (flipping) and affine transformation (translation) (**Figures 8**, **9**). Most neural networks include some component that makes it translationally-variant, such as a pooling layer or non-unary convolutional stride. In other words, the whole image is not necessarily the sum of individual tiles. In **Figure 7**, we demonstrate this effect due to a 2 × 2 max-pooling layer. Both the top and bottom use identical 11 × 2 arrays. If a 10 × 2 tile is used to perform the max-pooling, there are only two possible tiles. Notice that each tile produces different results. We further found that this behavior caused by the pooling layers most prominently affects the sharp intensity changes in object

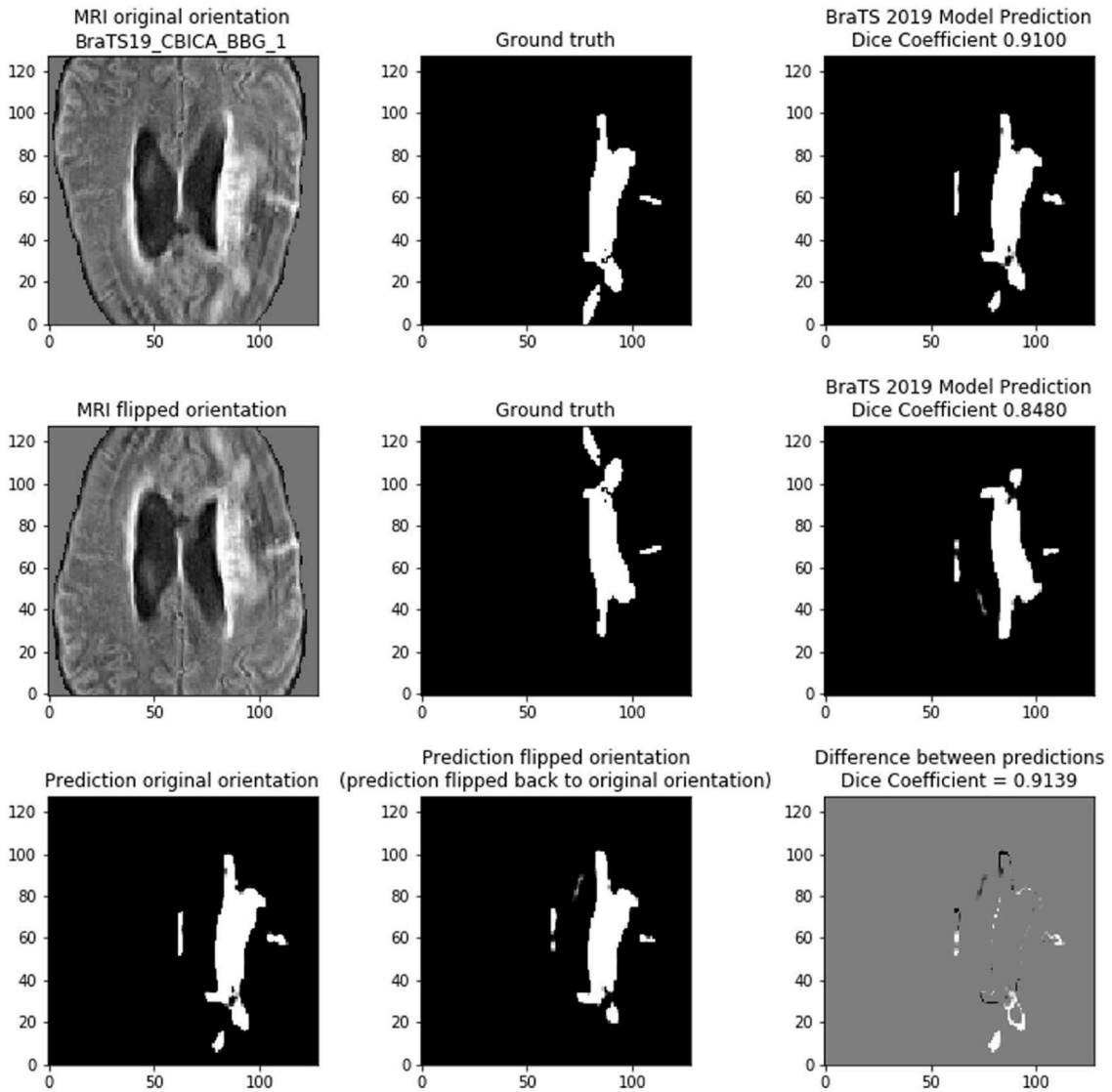


FIGURE 8 | Demonstrating the variability of the 2D BraTS model. **(Top)** Prediction based on normal orientation of the MRI input. **(Middle)** Prediction based on vertical flip of the MRI input. **(Bottom)** Comparing the predictions of the normal and flipped inputs. The prediction of the flipped input was re-flipped to allow direct comparison with the normal orientation prediction. In bottom right figure, gray pixels indicate no difference, black pixels are in the flipped prediction but are not present in the normal prediction, and white pixels are in the normal prediction but are not present in the flipped prediction.

boundaries. We believe that many of our results on "blobbier" borders that are more sensitive to even minor affine transforms to the tiles are a result of these translationally-variant operations, especially the max-pooling operation.

Although these differences in prediction are often localized to the segmentation border, the boundaries of the tumor or buildings are often the most relevant to the task. Especially in medical imaging, ensuring adequate tumor margins are critical to successful therapeutic planning and treatment.

4.1. Medical Data (BraTS)

If the models were linear, then any linear transformation to the model input should result in the same prediction (with the same linear transformation). **Figure 8** shows that on scan *BRATS19_CBICA_BBG_1.nii.gz* it achieves a *Dice* of 0.9100 on the center 128 × 128 tile of slice 94. However, if the MRI input is simply flipped vertically, then the prediction is changed. In this case, the *Dice* shows that the model provides a worse prediction with the flipped input (*Dice* = 0.8480). By reversing the linear transform (i.e., unflip the prediction) the two model predictions can be compared directly to show that they are indeed different (cross-prediction *Dice* = 0.9139). Although the two predictions are very similar, the bottom row of **Figure 8** highlights the differences occur along the tumor borders. We find that the tumor borders appear to be where the predictions differ.

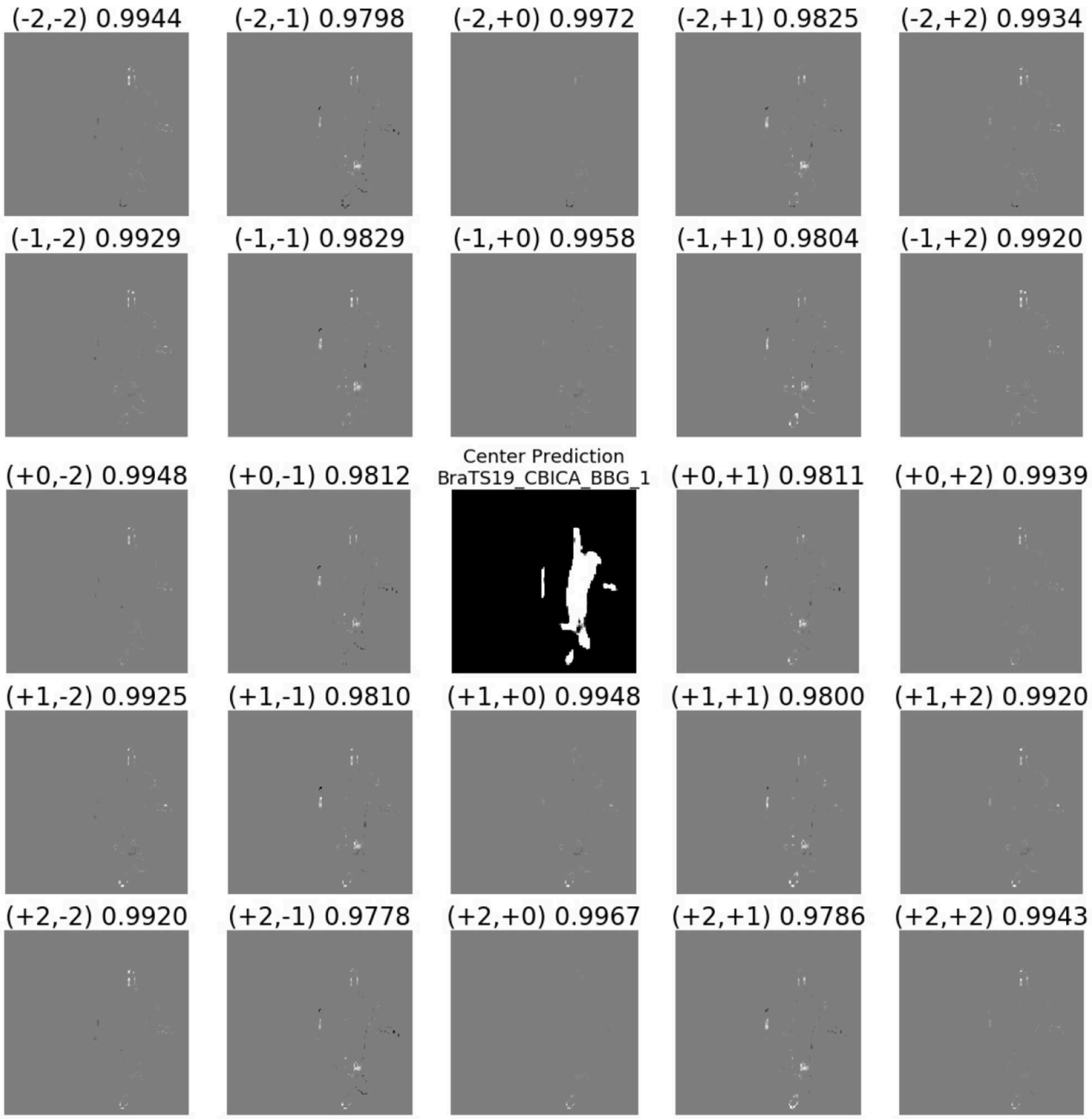


FIGURE 9 | Demonstrating the translational variance of the 2D BraTS model. The center shows the prediction of the model on a 128 × 128 center crop of the MRI. Note the entire tumor fits within this tile. The surrounding subplots show the difference in the prediction as the MRI input crop is offset by one or two pixels in either dimension of the slice. The differences between the predictions in the overlapping pixels show that even the smallest translation in the input can create a difference in the output. Gray pixels indicate no difference in the predictions of the overlapping pixels between the center crop and the offset crop, black pixels are in the offset-crop prediction but are not present in the center-crop prediction, and white pixels are in the center-crop prediction but are not present in the offset-crop prediction.

Figure 9 shows the translational variance of the model. The center shows the prediction of the model on a 128 × 128 center crop of the MRI. As the grid in the figure indicates, each tile shows the difference between pixels that overlap between the predictions of the center crop and a crop translated ±1 or 2 pixels in each dimension from the center crop. The *Dice* confirm that the overlapping predictions, while similar, differ significantly along the border of the tumor. This pattern of differences along the segmentation border was typical in the results. Note that the translations of (+2, +2) and (−2, −2) are a multiple of the max pooling stride and should be less sensitive to the translation (cf. Huang et al., 2018).

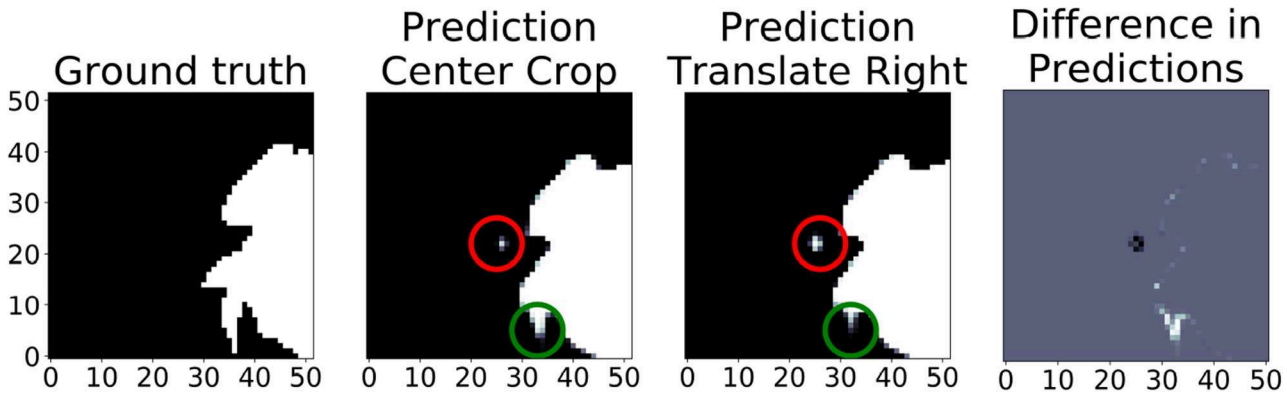


FIGURE 10 | In this "no pad BraTS" model, there is still translational variance despite the model containing no zero padding in the convolutional layers. The "Prediction Center Crop" refers to the prediction when using a 236 × 236 center crop of the input slice. "Prediction Translate Right" refers to a similar crop but translated one pixel to the right of the center. The green and red circles highlight predictions that have changed due translating by a single pixel. The figure on the right shows the difference in the overlapping regions between the "Center Crop" and "Translate Right" predictions. Gray indicates no difference. White indicates a prediction in the "Center Crop" that was not in the "Translate Right." Black indicates a prediction in the "Translate Right" that was not in the "Center Crop".

Figure 10 shows the translational variance of the "no pad BraTS" model. In this case, the model was trained without a zero pad in the convolutional layers so that we could assess the effects of zero padding on the prediction output. In the figure, the "Center Crop" refers to a 236 × 236 center crop of the 240 × 240 slice and "Prediction Translate Right" refers to a crop that has been translated one pixel to the right of the center crop. When we compare the prediction regions that overlap, we find several areas where the tumor prediction has changed (red and green circles). This demonstrates that the translational invariance due to tiling cannot be mitigated by simply modifying the topology to only use valid pixels in the convolutional layers.

Application of different tiling aggregation approaches (i.e., *rounding after averaging*, and *rounding before averaging*) revealed unpredictable and inconsistent results. This introduces a new parameter to standardize the results. The user must be aware of this discrepancy and make appropriate conclusion by experimenting with different tiling aggregation methods. Furthermore, the results of the 3D U-Net model inference demonstrate that greater image context (3D vs. 2D) contributes in the performance, but also that after the inclusion of sufficient image context (i.e., when providing enough context) the model converges and no further improvements are observed.

The two different tiling aggregation approaches produced different results in the 2D and 3D models. For the 2D model the *rounding after averaging* approach produced a substantially lower *Dice* metric than *rounding before averaging* approach. In the 3D model, the *rounding after averaging* approach produced an insignificantly higher *Dice* metric than *rounding before averaging*.

4.2. Non-medical Data (SpaceNet-Vegas)

We find that the whole image consistently outperforms tiling-based approaches on the pixelwise *Dice* that converge to similar values once the tile size reached approximately half of the original height and width of the image (**Table 3**). Similarly, the polygonal-wise *Dice* (SpaceNet *F*1 metric) also improves as the training tiles cover a larger proportion of the whole image. Inspection of the predicted masks reveals the likely culprit: **Figure 11** shows the ground truth mask and image at the top, followed by rows showing sliding window predictions with 128 × 128 and 256 × 256 tiles, with the last row being predictions from the model trained on 640 × 640 resized inputs.

We note that predictions from the models using smaller tile sizes produce segmentations that fail to capture fine-grained boundaries between buildings, leading to "blobbier" or more amorphous predictions. Note that the tiled predictions segment the buildings in each cul-de-sac as a single continuous mass; however, there are roughly 6 houses per cul-de-sac. These missing boundaries lead to the post-processing step of polygonization creating a reduced number of polygons as multiple buildings are getting extracted as one. We note that as the tile size increases to reach at least $\frac{1}{2}H \times \frac{1}{2}W$, then the adverse polygonization effects are reduced and predictions at the boundary of the segmentations becomes more accurate (**Table 3**).

The removal of zero padding in the topology has a negligible effect on the average per-image *Dice* coefficient. However, the *F*1 Metric is lower by 1–3% across all input size variants (**Table 4**). Because the removal of zero-padding reduces the output size of the model, what we see is an effect similar to the discussed effects of using smaller tiles rather than the whole image during training. Since the *F*1 metric is computed over the entirety of the extracted polygons—that is, a set latitude/longitude pairs defining a building footprint—we again lose fidelity at the edges of the buildings which decreased the SpaceNet *F*1 score. This does not affect the *Dice* score, as *Dice* is not sensitive to the separation of object instances. In other word, a giant pixel blob covering two buildings yields good pixelwise *Dice* values, but poor SpaceNet *F*1 polygon values.

5. CONCLUSIONS

In this study, we systematically evaluated the effects of using tiling approaches vs. using the whole image for deep learning semantic segmentation, in both 2D and 3D configurations.

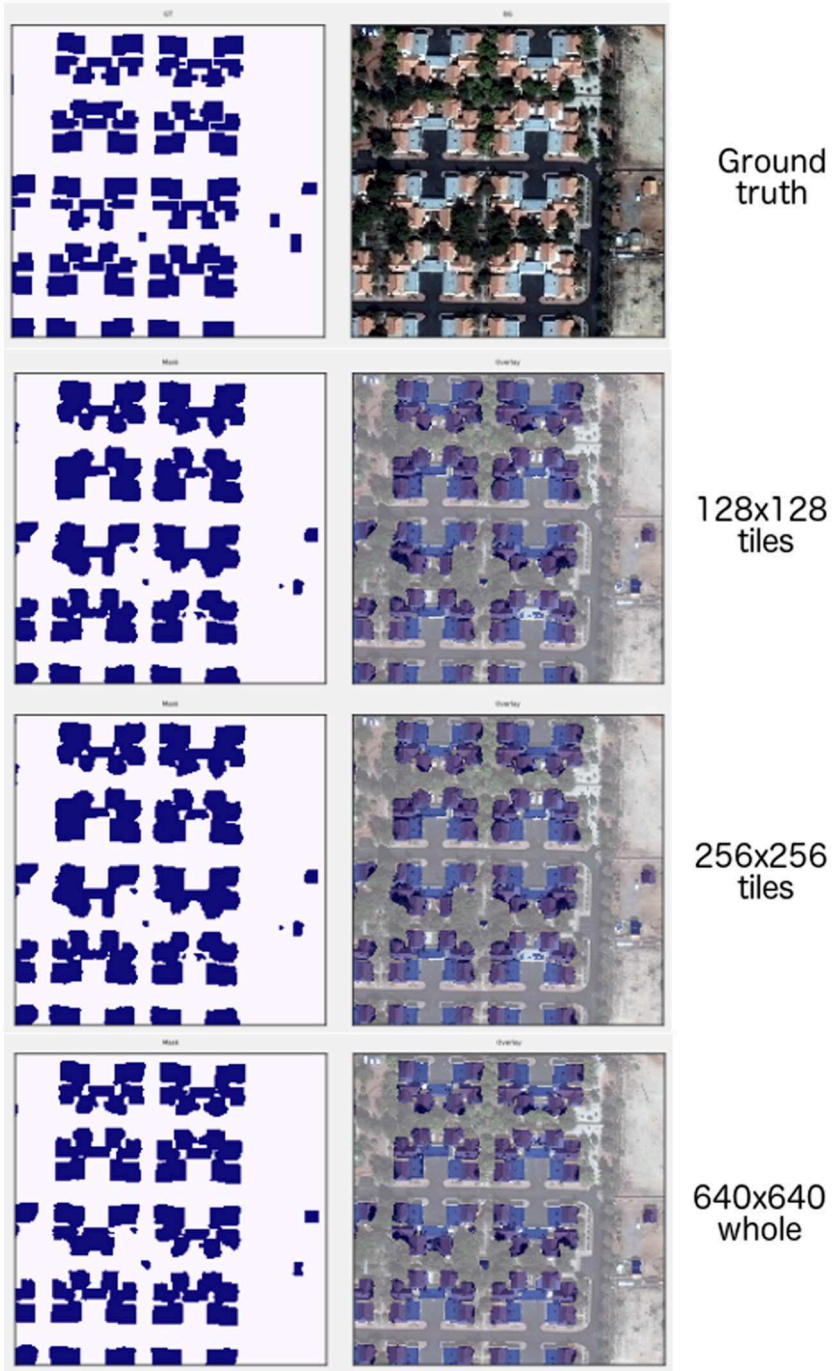


FIGURE 11 | (1st row) Ground truth and whole image. (2nd row) 128 × 128 tiles. (3rd row) 256 × 256 tiles. (4th row) 640 × 640 whole image.

Through quantitative evaluation we demonstrated that larger tile (i.e., context) sizes yield more consistent results and mitigate undesirable and unpredictable behavior during inference. We realize that tiling methods may continue to be necessary as researchers use images with increasingly greater size and resolution in their convolutional neural network models. Our goal in this study is to raise awareness about the issues surrounding tiling. Namely:

1. Tiling hyperparameters, which include tile size, offset, orientation, and overlap, can cause large variations in the prediction, particularly around the borders of the segmentation mask.
2. This variance is not just limited to a translation less than the stride (as suggested by Huang et al., 2018), but seem to be present even with translations of ±2 in each direction. Therefore, we think that our results show

a more complicated story to the translational variance of CNNs.
3. Topologies without zero padding in the convolutional layers do not eliminate the translational variance of the topology.
4. Methods to aggregate the individual predictions into a whole image prediction, namely when to average the predicted outcome pseudo-probability maps and when to round these predictions, that can have a significant effect on the overall accuracy.
5. Larger degrees of image context, including adding 3D information to the model and using larger tile sizes, improves model performance in training and is less sensitive to these hyperparameters during inference.

We conclude that increased access to memory—either through improvements in hardware or through high performance computing techniques, such as model parallelism (Shazeer et al., 2018) and data parallelism (Sergeev and Balso, 2018)—is essential to creating accurate and robust models. Tiling should only be reserved for those cases where the physical limitations of memory make it an absolute necessity. When tiling must be used, researchers should be careful to investigate how the translational variance of the model affects the predictions and compare methods of tiling aggregation to determine the best way to mitigate the variability inherent in tiling.

DATA AVAILABILITY STATEMENT

The datasets used for this study are publicly available as parts of the International Brain Tumor Segmentation challenge 2019 (https://www.med.upenn.edu/cbica/brats2019.html) and the SpaceNet satellite imagery dataset (https://spacenet.ai/spacenet-buildings-dataset-v2/). Appropriate citations for each of the datasets are provided within the article.

AUTHOR CONTRIBUTIONS

GR, RP, and SB: study conception and design. GR, RP, ST, and AB: software development used in the study. GR, RP, AB, and SB: wrote the paper. GR, RP, ST, AB, and SB: data analysis and interpretation, and reviewed/edited the paper.

FUNDING

Research reported in this publication was partly supported by the National Institutes of Health (NIH) under award numbers NINDS:R01NS042645, NCI:U24CA189523, NCI:U01CA242871. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH.

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM Collection*. The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG Collection*. The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* 1811.02629.

Chen, T., Xu, B., Zhang, C., and Guestrin, C. (2016). Training deep nets with sublinear memory cost. *arXiv:1604.06174*. Available online at: http://arxiv.org/abs/1604.06174

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer International Publishing), 424–432.

Ghosh, A., Ehrlich, M., Shah, S., Davis, L. S., and Chellappa, R. (2018). "Stacked U-Nets for ground material segmentation in remote sensing imagery," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Salt Lake City, UT), 252–2524. doi: 10.1109/CVPRW.2018.00047

Greenspan, H., Van Ginneken, B., and Summers, R. M. (2016). Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans. Med. Imaging* 35, 1153–1159. doi: 10.1109/TMI.2016.2553401

Hagerty, P. (2016). *The SpaceNet Metric*. Available online at: https://medium.com/the-downlinq/the-spacenet-metric-612183cc2ddb (accessed April 22, 2019).

Huang, B., Reichman, D., Collins, L. M., Bradbury, K., and Malof, J. M. (2018). *Tiling and Stitching Segmentation Output for Remote Sensing: Basic Challenges and Recommendations. arXiv:1805.12219*. Available online at: http://arxiv.org/abs/1805.12219

Isensee, F., Petersen, J., Kohl, S. A. A., Jäger, P. F., and Maier-Hein, K. H. (2019). *NNU-Net: Breaking the Spell on Successful Medical Image Segmentation. arXiv:1904.08128*. Availble online at: http://arxiv.org/abs/1904.08128

Ito, Y., Imai, H., Duc, T. L., Negishi, Y., Kawachiya, K., Matsumiya, R., et al. (2019). Profiling based out-of-core hybrid method for large neural networks. *arXiv:1907.05013*.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc.), 1097–1105. doi: 10.1145/3065386

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (BraTS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Pinckaers, H., and Litjens, G. (2018). Training convolutional neural networks with megapixel images. *arXiv:1804.05712*. Available online at: http://arxiv.org/abs/1804.05712

Reina, G. A., and Panchumarthy, R. (2018). "Adverse effects of image tiling on convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Cham: Springer International Publishing), 25–36.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer International Publishing), 234–241.

Roth, H. R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., et al. (2018). An application of cascaded 3D fully convolutional networks for medical image segmentation. *Comput. Med. Imaging Graph.* 66, 90–99. doi: 10.1016/j.compmedimag.2018.03.001

Sergeev, A., and Balso, M. D. (2018). Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv:1802.05799*. Available online at: http://arxiv.org/abs/1802.05799

Shazeer, N., Cheng, Y., Parmar, N., Tran, D., Vaswani, A., Koanantakool, P., et al. (2018). "Mesh-TensorFlow: deep learning for supercomputers," in *Neural Information Processing Systems* (Montreal, QC).

SpaceNet (2018). *SpaceNet on Amazon Web Services (AWS) Datasets*. The SpaceNet Catalog. Available online at: https://spacenetchallenge.github.io/datasets/datasetHomePage.html (accessed April 22, 2019).

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. *arXiv:1607.08022*. Available online at: http://arxiv.org/abs/1607.08022

Weir, N., Lindenbaum, D., Bastidas, A., Etten, A. V., McPherson, S., Shermeyer, J., et al. (2019). Spacenet MVOI: a multi-view overhead imagery dataset. *CoRR* abs/1903.12239. Available online at: http://arxiv.org/abs/1903.12239

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv:1505.00853*. Available online at: http://arxiv.org/abs/1505.00853

Zeng, G., and Zheng, G. (2018). Holistic decomposition convolution for effective semantic segmentation of 3D MR images. *arXiv:1812.09834*. Available online at: http://arxiv.org/abs/1812.09834

Conflict of Interest: GR, RP, and AB were employed by the company Intel Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 14 February 2020
doi: 10.3389/fncom.2020.00009

Segmenting Brain Tumor Using Cascaded V-Nets in Multimodal MR Images

Rui Hua[1,2], Quan Huo[2], Yaozong Gao[2], He Sui[3], Bing Zhang[4], Yu Sun[1], Zhanhao Mo[3] and Feng Shi[2]**

[1] School of Biological Science and Medical Engineering, Southeast University, Nanjing, China, [2] Shanghai United Imaging Intelligence, Co., Ltd., Shanghai, China, [3] China-Japan Union Hospital of Jilin University, Changchun, China, [4] Department of Radiology, Affiliated Drum Tower Hospital of Nanjing University Medical School, Nanjing, China

In this work, we propose a novel cascaded V-Nets method to segment brain tumor substructures in multimodal brain magnetic resonance imaging. Although V-Net has been successfully used in many segmentation tasks, we demonstrate that its performance could be further enhanced by using a cascaded structure and ensemble strategy. Briefly, our baseline V-Net consists of four levels with encoding and decoding paths and intra- and inter-path skip connections. Focal loss is chosen to improve performance on hard samples as well as balance the positive and negative samples. We further propose three preprocessing pipelines for multimodal magnetic resonance images to train different models. By ensembling the segmentation probability maps obtained from these models, segmentation result is further improved. In other hand, we propose to segment the whole tumor first, and then divide it into tumor necrosis, edema, and enhancing tumor. Experimental results on BraTS 2018 online validation set achieve average Dice scores of 0.9048, 0.8364, and 0.7748 for whole tumor, tumor core and enhancing tumor, respectively. The corresponding values for BraTS 2018 online testing set are 0.8761, 0.7953, and 0.7364, respectively. We also evaluate the proposed method in two additional data sets from local hospitals comprising of 28 and 28 subjects, and the best results are 0.8635, 0.8036, and 0.7217, respectively. We further make a prediction of patient overall survival by ensembling multiple classifiers for long, mid and short groups, and achieve accuracy of 0.519, mean square error of 367240 and Spearman correlation coefficient of 0.168 for BraTS 2018 online testing set.

Keywords: deep learning, brain tumor, segmentation, V-Net, multimodal, magnetic resonance imaging

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Eranga Ukwatta,
Johns Hopkins University,
United States
Siddhesh Pravin Thakur,
University of Pennsylvania,
United States
Anahita Fathi Kazerooni,
University of Pennsylvania,
United States

***Correspondence:**
Zhanhao Mo
mozhanhao@jlu.edu.cn
Feng Shi
feng.shi@united-imaging.com

Received: *28 April 2019*
Accepted: *24 January 2020*
Published: *14 February 2020*

Citation:
Hua R, Huo Q, Gao Y, Sui H, Zhang B, Sun Y, Mo Z and Shi F (2020) Segmenting Brain Tumor Using Cascaded V-Nets in Multimodal MR Images. Front. Comput. Neurosci. 14:9. doi: 10.3389/fncom.2020.00009

INTRODUCTION

Gliomas are the most common brain tumors and comprise about 30 percent of all brain tumors. Gliomas occur in the glial cells of the brain or the spine (Mamelak and Jacoby, 2007). They can be further categorized into low-grade gliomas (LGG) and high-grade gliomas (HGG) according to their pathologic evaluation. LGG are well-differentiated and tend to exhibit benign tendencies and portend a better prognosis for the patients. HGG are undifferentiated and tend to exhibit malignant and usually lead to a worse prognosis. With the development of the magnetic resonance imaging (MRI), multimodal MRI plays an important role in disease diagnosis. Different MRI

modalities are sensitive to different tumor tissues. For example, T2-weighted (T2) and T2 Fluid Attenuation Inversion Recovery (FLAIR) are sensitive to peritumoral edema, and post-contrast T1-weighted (T1Gd) is sensitive to necrotic core and enhancing tumor core. Thus, they can provide complementary information about gliomas.

Segmentation of brain tumor is a prerequisite while essential task in disease diagnosis, surgical planning and prognosis (Bakas et al., 2017a). Automatic segmentation provides quantitative information that is more accurate and has better reproducibility than conventional qualitative image review. Moreover, the following task of brain tumor classification heavily relies on the results of brain tumor segmentation. Automatic segmentation is considered as a powered engine and empower other intelligent medical application. However, the segmentation of brain tumor in multimodal MRI scans is one of the most challenging tasks in medical imaging analysis due to their highly heterogeneous appearance, and variable localization, shape and size.

Before deep learning developed, random forest (RF) achieves better performance in brain tumor segmentation (Zikic et al., 2012; Le Folgoc et al., 2016). In recent years, with the rapid development of deep leaning techniques, state-of-the-art performance on brain tumor segmentation have been achieved with convolutional neural network (CNN). For example, in Cui et al. (2018), an end-to-end training using fully convolutional network (FCN) showed satisfactory performance in the localization of the tumor, and patch-wise CNN was used to segment the intra-tumor structure. In Wang et al. (2018), a cascaded anisotropic CNN was designed to segment three sub-regions with three Nets, and the segmentation result from previous net was used as receptive field in the next net. Ensemble strategy also shows great advantages, and most models are based on 3D U-Net, DeepMedic, and their variants (Isensee et al., 2018; Kamnitsas et al., 2018). One recent paper arguing that a well-trained U-Net is hard to beat (Isensee et al., 2019). Instead of modifying architectures, they focused on the training process such as region based training and additional training data, and achieved competitive Dice scores.

Inspired by the superior performance of V-Net in segmentation tasks, we propose a cascaded V-Nets method to segment brain tumor into three substructures and background. In particular, the cascaded V-Nets not only take advantage of residual connection but also use the extra coarse localization and ensemble of multiple models to boost the performance. A preliminary version of the method has been presented in a conference (Hua et al., 2019). Here we extend it to include more descriptions of the method details and additional experiments to further evaluate the performance of the proposed method in local hospital data sets.

METHOD

Dataset and Preprocessing

The data used in experiments come from the released data of BraTS 2018 online challenge (Menze et al., 2015; Bakas et al., 2017a,b,c). The training set includes totally 210 HGG patients and 75 LGG patients. The validation set includes 66 patients and the testing set includes 191 patients. Each patient has four MRI modalities including T1-weighted (T1), T2, T1Gd, and FLAIR, where ground truth labels of tumor substructures are available only in training set. The images were already skull stripped and normalized together, with resolution of $1 \times 1 \times 1$ mm^3 for all modalities. We use 80 percent of the training data for our training, and the rest 20 percent of the training data as our local testing set.

Meanwhile, in order to further test the performance of the proposed method, we prepare two additional data sets that include 28 patients from China-Japan Union Hospital of Jilin University and another 28 patients from Affiliated Drum Tower Hospital of Nanjing University Medical School. The resolution of the T1 images from China-Japan Union Hospital of Jilin University is $0.6 \times 0.6 \times 6$ mm^3, while the resolution of the T1 images from Affiliated Drum Tower Hospital of Nanjing University Medical School is $0.67 \times 0.67 \times 0.67$ mm^3. The images of T2, T1Gd, and FLAIR are linearly aligned to its corresponding T1 image for each subject. Skull stripping is performed on T1 and the mask is applied to other modalities. The ground truth labels of the brain tumors are manually delineated by an experienced radiologist. The experienced radiologist (Z.M.) was asked to delineate the tumor subregions according to the image delineating principles of BraTS 2018. Results would serve as ground truth to evaluate the generalizability of the method. In detail, the delineating principle includes three subregion segmentations of the tumor, including the necrotic (NCR) and the non-enhancing (NET) tumor core, the enhancing tumor (ET) and the peritumoral edema (ED). The NCR and the NET tumor core was the low intensity necrotic structures in T1Gd when compared to T1. The ET area was confirmed as hyper-intensity structures in T1Gd when compared to T1 images, and when compared to normal brain in T1Gd. The ED area was identified as abnormality visible in T2 and FLAIR excluding ventricles and cerebrospinal fluid.

All data used in the experiments are preprocessed with specific designed procedures. A flow chart of the proposed preprocessing procedures is shown in **Figure 1**, as follows: (1) Apply bias field correction N4 (Tustison et al., 2010) to T1 and T1Gd images, normalize each modality using histogram matching with respect to a MNI template image, and rescale the images intensity values into range of -1 to 1; (2) Apply bias field correction N4 to all modalities, compute the standardized z-scores for each image and rescale 0–99.9 percentile intensity values into range of -1 to 1; (3) Follow the first method, and further apply affine alignment to co-register each image to the MNI template image.

V-Net Architecture

V-Net was initially proposed to segment prostate by training an end-to-end CNN on MRI (Milletari et al., 2016). The architecture of our V-Net is shown in **Figure 2**. The left side of V-Net reduces the size of the input by down-sampling, while the right side of V-Net recovers the semantic segmentation image that has the same size with input images by applying de-convolutions. The detailed parameters about V-Net is shown in **Table 1**. Both left side of the network and right side of the network were divided into four blocks that operate at different resolutions. Each block comprises

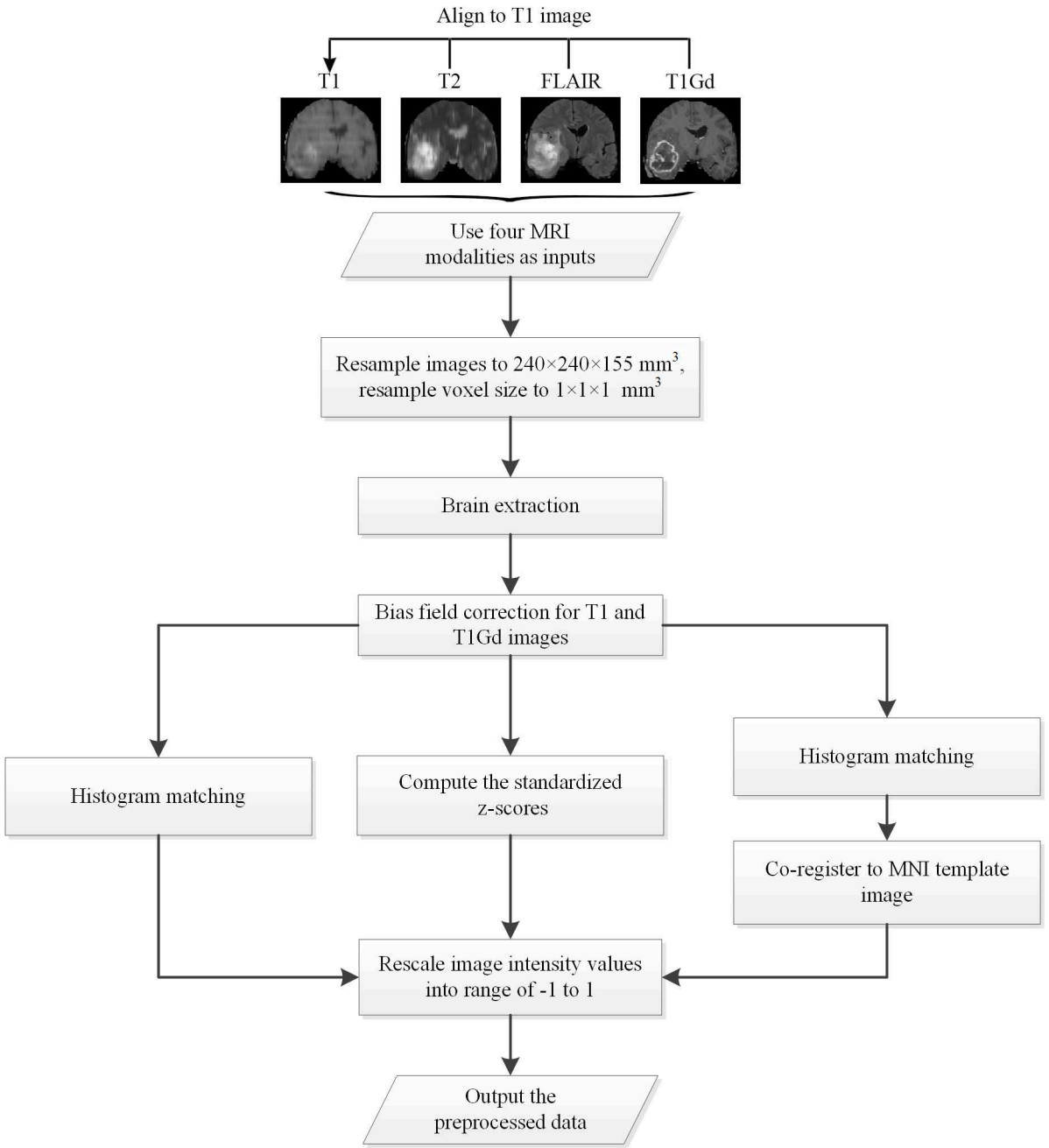


FIGURE 1 | The flow chart of the preprocessing procedures.

one to three convolutional blocks. The input of each block is added to the output of the current block to learn a residual function, and added to the input of the corresponding block which has the same resolution in the right side of the network as a skip connection. By means of introducing residual function and skip connection, V-Net has better segmentation performance compared with conventional CNN. Each convolutional block comprises two convolutional layers with the kernel size of $1 \times 1 \times 1$ at the start and the end of the convolutional block. By means of introducing the 3D kernel with size of $1 \times 1 \times 1$, the number of parameters in V-Net is decreased and the memory consumption is greatly reduced. Appropriate padding and ReLU non-linearity are applied throughout the network.

Proposed Cascaded V-Nets Framework

Although V-Net has demonstrated promising performance in segmentation tasks, it could be further improved if incorporated with extra information, such as coarse localization. Therefore,

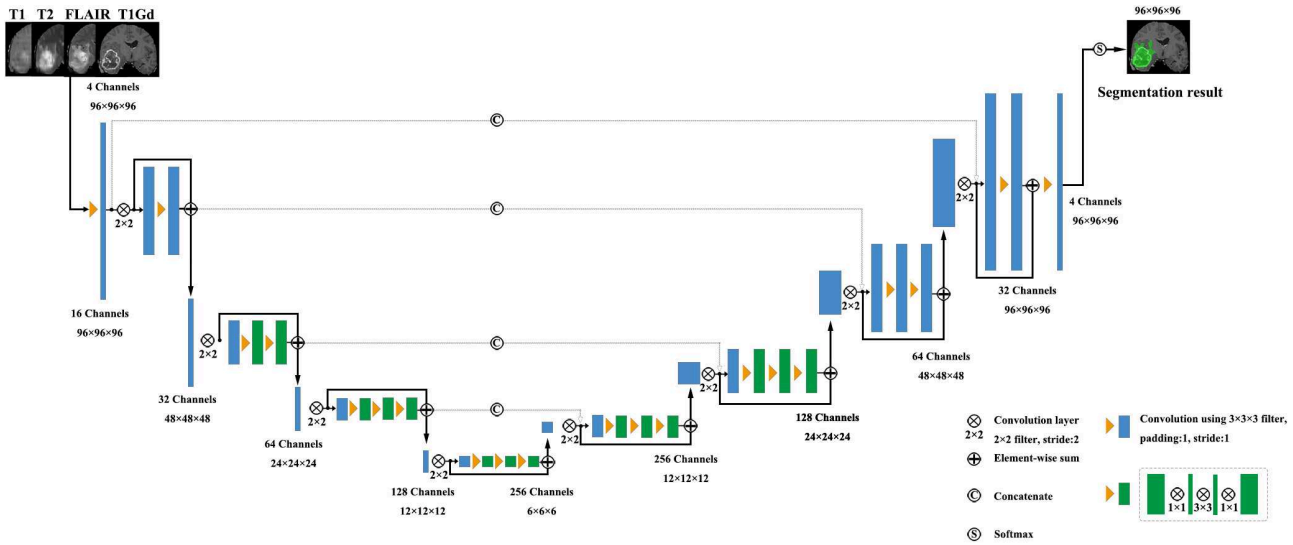


FIGURE 2 | The architecture of the used V-Net.

we propose a cascaded V-Nets method for tumor segmentation. Briefly, we (1) use one V-Net for the whole tumor segmentation; (2) use a second V-Net to further divide the tumor regions into three substructures, e.g., tumor necrosis, edema, and enhancing tumor. Note that the coarse segmentation of whole tumor in the first V-Net is also used as receptive field to boost the performance. Detailed steps are as follows.

The proposed framework is shown in **Figure 3**. There are two networks to segment substructures of brain tumors sequentially. The first network (V-Net 1) includes models 1–3, designed to segment the whole tumor. These models are trained by three kinds of preprocessed data mentioned in part of 2.1, respectively. V-Net 1 uses four modalities MR images as inputs, and outputs the mask of whole tumor (WT). The second network (V-Net 2) includes models 4–5, designed to segment the brain tumor into three substructures: tumor necrosis, edema, and enhancing tumor. These models are trained by the first two kinds of preprocessed data mentioned in part of 2.1, respectively. V-Net 2 also uses four modalities MR images as inputs, and outputs the segmented mask with three labels. Note that the inputs of V-Net 2 have been processed using the mask of WT as region of interest (ROI). In other words, the areas out of the ROI are set as background. Finally, we combine the segmentation results of whole tumor obtained by V-Net 1 and the segmentation results of tumor core (TC, includes tumor necrosis and enhancing tumor) obtained by V-Net 2 to achieve more accurate results about the three substructures of brain tumor. In short, the cascaded V-Nets take advantage of segmenting the brain tumor and three substructures sequentially, and ensemble of multiple models to boost the performance and achieve more accurate segmentation results.

Ensemble Strategy

We employ a simple yet efficient ensemble strategy. It works by averaging the probability maps obtained from different models. We use ensemble strategy twice in the two-step segmentation of the brain tumor substructures. For example, in V-Net 1, the probability maps of WT obtained from model 1, model 2, and model 3 were averaged to get the final probability map of WT. In V-Net 2, the probability maps of tumor necrosis, edema, and enhancing tumor obtained from model 4 and model 5 were averaged to get final probability maps of brain tumor substructures, respectively. In order to evaluate the effect of ensemble strategy for enhancing the performance of our cascaded V-Nets, ablation experiments were conducted on MICCAI BraTS 2018 validation dataset. Briefly, model combinations include Model 1–4, Model 12–4, Model 123–4, Model 123–45, and Model 123–45-fuse. To evaluate the significance of the results between different model combinations, we first evaluated the overall difference across model combinations with Kruskal-Wallis H test, and then checked the difference between each of two groups with Mann-Whitney U test. Multiple comparison correction was performed using Bonferroni criteria.

Network Implementation

Our cascaded V-Nets are implemented in the deep learning framework PyTorch. In our network, we initialize weights with kaiming initialization (He et al., 2015), and use focal loss (Lin et al., 2018) illustrated in formula (1) as loss function. Focal loss has the advantage of balancing the ratio of positive and negative samples, and decreases the importance of easy classified samples to focus more on difficult samples (Lin et al., 2018). Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014) is used as optimizer with learning rate of 0.001, and batch size of 8. Experiments are performed with a NVIDIA Titan Xp 12GB GPU.

$$Focal\_Loss\left(p_t\right) = -\alpha\left(\mathbf{1}-p_t\right)^r \mathbf{log}\left(p_t\right) \tag{1}$$

where, α denotes the weight to balance the importance of positive/negative samples, r denotes the factor to

TABLE 1 | The detailed parameters of the used V-Net, as shown in **Figure 2**.

Blocks	Sub-blocks or layers	Input dimensions	Output dimensions
Input block	Conv(k = 3, p = 1, s = 1) + BN + ReLU	96 × 96 × 96 × 4	96 × 96 × 96 × 16
Down block 1	Conv(k = 2, p = 0, s = 2)+ BN + ReLU	96 × 96 × 96 × 16	48 × 48 × 48 × 32
	Conv(k = 3, p = 1, s = 1) + BN*	48 × 48 × 48 × 32	–
	(input+output) + ReLU*	48 × 48 × 48 × 32	–
Down block 2	Conv(k = 2, p = 0, s = 2) + BN + ReLU	48 × 48 × 48 × 32	24 × 24 × 24 × 64
	Conv block × 2*	24 × 24 × 24 × 64	–
	(input+output) + ReLU*	24 × 24 × 24 × 64	–
Down block 3	Conv(k = 2, p = 0, s = 2) + BN + ReLU	24 × 24 × 24 × 64	12 × 12 × 12 × 128
	Conv block × 3*	12 × 12 × 12 × 128	–
	(input+output) + ReLU*	12 × 12 × 12 × 128	–
Down block4	Conv(k = 2, p = 0, s = 2) + BN + ReLU	12 × 12 × 12 × 128	6 × 6 × 6 × 256
	Conv block × 3*	6 × 6 × 6 × 256	–
	(input+output) + ReLU*	6 × 6 × 6 × 256	–
Up block 1	Conv(k = 2, p = 0, s = 2) + BN + ReLU	6 × 6 × 6 × 256	12 × 12 × 12 × 128
	Cat(output, skip)*	12 × 12 × 12 × 128	12 × 12 × 12 × 256
	Conv block × 3*	12 × 12 × 12 × 256	–
	(input+output) + ReLU*	12 × 12 × 12 × 256	–
Up block 2	Conv(k = 2, p = 0, s = 2) + BN + ReLU	12 × 12 × 12 × 256	24 × 24 × 24 × 64
	Cat(output+skip)*	24 × 24 × 24 × 64	24 × 24 × 24 × 128
	Conv Block × 3*	24 × 24 × 24 × 128	–
	(input+output) + ReLU*	24 × 24 × 24 × 128	–
Up block 3	Conv(k = 2, p = 0, s = 2) + BN + ReLU	24 × 24 × 24 × 128	48 × 48 × 48 × 32
	Cat(output+skip)*	48 × 48 × 48 × 32	48 × 48 × 48 × 64
	Conv(k = 3, p = 1, s = 1) + BN + ReLU*	48 × 48 × 48 × 64	–
	Conv(k = 3, p = 1, s = 1)+BN*	48 × 48 × 48 × 64	–
	(input+output) + ReLU*	48 × 48 × 48 × 64	–
Up block 4	Conv(k = 2, p = 0, s = 2) + BN + ReLU	48 × 48 × 48 × 64	96 × 96 × 96 × 16
	Cat(output+skip)*	96 × 96 × 96 × 16	96 × 96 × 96 × 32
	Conv(k = 3, p = 1, s = 1) + BN*	96 × 96 × 96 × 32	–
	(input+output) + ReLU*	96 × 96 × 96 × 32	–
Out block	Conv(k = 1, p = 0, s = 1) + BN + ReLU	96 × 96 × 96 × 32	96 × 96 × 96 × 4
	Softmax	96 × 96 × 96 × 4	96 × 96 × 96 × 1

Each Conv sub-block contains three convolution layers: Conv1 (k = 1, p = 0, s = 1), Conv2 (k = 3, p = 1, s = 1), and Conv3 (k = 1, p = 0, s = 1). k, kernel size; p, padding; s, stride. The symbol "–" means the output dimensions are the same with input dimensions. The symbol "" denotes that these layers in each block are residual units.*

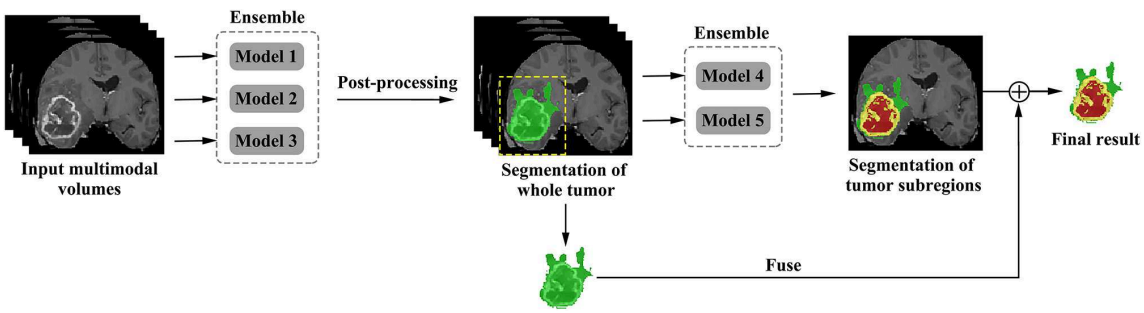


FIGURE 3 | The proposed framework of cascaded V-Nets for brain tumor segmentation.

increase the importance of correcting misclassified samples, and p_t denotes the probability of the ground truth.

In order to reduce the memory consumption in the training process, 3D patches with a size of 96 × 96 × 96 are used. And the center of the patch is confined to the bounding box of the

brain tumor. Therefore, every patch used in training process contains both tumor and background. The training efficiency of the network has been greatly improved.

Post-processing

The predicted tumor segmentations are post-processed using connected component analysis. We consider that the isolated segmentation labels with small size are prone to artifacts and thus remove them. Our strategy is as follows. After the V-Net 1, the small clusters with voxel number <T = 1,000 are directly discarded. For each cluster with size between 1,000 and 15,000, its average probability of being a tumor is calculated. This cluster will be retained if the probability is no <0.85 and removed otherwise. The rest big clusters with voxel number over T = 15,000 are also retained. A binary whole tumor map is thus obtained. After the V-Net 2, we also calculated the connected component and removed the small clusters with voxel number <1,000. While if all cluster sizes are <1,000, the largest cluster will be retained.

Evaluation of Tumor Segmentation Performance

The models trained by MICCAI BraTS 2018 training data are applied to our local testing set, MICCAI BraTS 2018 validation set, MICCAI BraTS 2018 testing set, and the additional clinical testing sets. In order to evaluate the performance of our method, Dice score, sensitivity, and specificity are calculated for whole tumor, tumor core and enhancing tumor, respectively. Dice score indicates the ratio of the area where the segmentation image intersects with the ground truth image to the total areas. Sensitivity indicates the ratio of the detected tumor voxels to all tumor voxels. Specificity indicates the ratio of the detected background voxels to all background voxels. The evaluation results for MICCAI BraTS 2018 validation set and testing set are provided by the organizer of the BraTS 2018 online challenge, and Hausdorff95 is also included, which indicates the distances of the two tumor voxels sets with a percentile value of 95%.

$$Dice = \frac{2\left|A \bigcap B\right|}{|A| + |B|} \tag{2}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4}$$

$$Hausdorff95 = max[\underset{a\in A}{max}(95\%)\underset{b\in B}{min}\|\boldsymbol{a}\text{ - }\boldsymbol{b}\|, \underset{b\in B}{max}(95\%)\underset{a\in A}{min}\left\|b - a\right\|] \tag{5}$$

where, A denotes the segmentation image, B denotes the ground truth image, TP denotes the number of the true positive voxels, FN denotes the number of the false negative voxels, TN denotes the number of the true negative voxels, and FP denotes the number of the false positive voxels.

For the additional testing sets of local hospitals, only Dice scores are evaluated. Given that the images from two data sets have different resolution, we calculate the average Dice scores for whole tumor, tumor core and enhancing tumor in two data sets, respectively.

TABLE 2 | Selected features in the training data for the prediction of patient overall survival.

Features	Number of features
Age	1
Volume of whole brain	1
Volume of whole tumor	1
Volumes of three tumor substructures	3
Ratio of the whole tumor in whole brain	1
Ratios of three tumor substructures in whole tumor	3
Extent of lesion in x, y, z directions	3
Center coordinates of the whole tumor	3
Means and variances of three tumor substructures in four MR modalities	24
First order statistics features of three tumor substructures	411
Shape-based features of three tumor substructures	78
Gray level cooccurence matrix features of three tumor substructures	180
Gray level run length matrix features of three tumor substructures	96
Neigbouring gray tone difference matrix features of three tumor substructures	96
Gray level dependence matrix features of three tumor substructures	84

Prediction of Patient Overall Survival

Overall survival (OS) is a direct measure of clinical benefit to a patient. Generally, brain tumor patients could be classified into long-survivors (e.g., >15 months), mid-survivors (e.g., between 10 and 15 months), and short-survivors (e.g., <10 months). For the multimodal MRI data, we propose to use our tumor segmentation masks and generate imaging markers through Radiomics method to predict the patient OS groups.

From the training data, we extract 40 hand-crafted features and 945 radiomics features (Isensee et al., 2018) in total. The detailed extracted features are shown in **Table 2**. All features are normalized into range of 0–1. Pearson correlation coefficient is used for feature selection. All features are ranked by Pearson correlation coefficient from large to small, and the top 10% features are used as the inputs of the following classifiers. We use support vector machine (SVM), multilayer perceptrons (MLP), XGBoost, decision tree classifier, linear discriminant analysis (LDA), and random forest (RF) as our classifiers in an ensemble strategy. F1-score is used as the evaluation standard. The final result is determined by the vote on all classification results. In order to reduce the bias, a 10-fold cross-validation is used. For the validation and testing data, these selected features are extracted and the prediction is made using the above models.

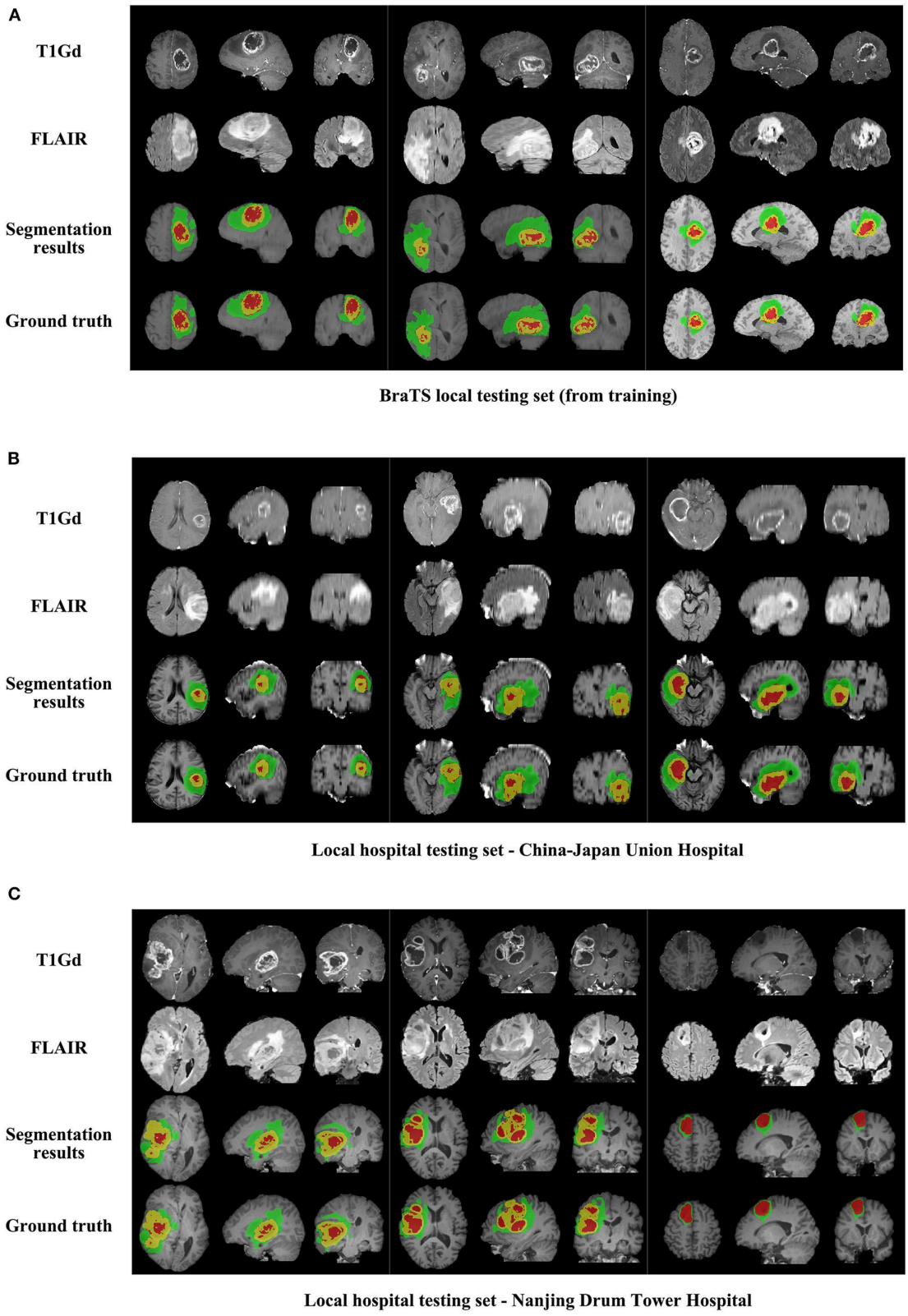


FIGURE 4 | The comparison of segmentation results and ground truth on representative cases from local testing set and two clinical testing sets. **(A)** The segmentation results and ground truth from local testing set. **(B)** The segmentation results and ground truth from clinical testing set of China-Japan Union Hospital of Jilin University. **(C)** The segmentation results and ground truth from clinical testing set of Affiliated Drum Tower Hospital of Nanjing University Medical School.

TABLE 3 | Dice, sensitivity, and specificity measurements of the proposed method on local testing set.

	Whole tumor	Tumor core	Enhancing tumor
Dice mean ± SD	0.8505 ± 0.0972	0.7842 ± 0.1919	0.7426 ± 0.2080
Sensitivity mean ± SD	0.9180 ± 0.1091	0.7596 ± 0.2199	0.7174 ± 0.2337
Specificity mean ± SD	0.9981 ± 0.0012	0.9996 ± 0.0008	0.9997 ± 0.0003

TABLE 4 | Dice, sensitivity, specificity, and Hausdorff95 measurements of the proposed method on BraTS 2018 validation set.

	Whole tumor	Tumor core	Enhancing tumor
Dice mean ± SD	0.9048 ± 0.0648	0.8364 ± 0.1609	0.7768 ± 0.2355
Sensitivity mean ± SD	0.9146 ± 0.0949	0.8453 ± 0.1781	0.8166 ± 0.2382
Specificity mean ± SD	0.9945 ± 0.0041	0.9971 ± 0.0041	0.9977 ± 0.0032
Hausdorff95 mean ± SD (mm)	5.1759 ± 7.3622	6.2780 ± 7.7681	3.5123 ± 4.5407

RESULTS

Segmentation Results on Local Testing Set of 57 Subjects

We use 20 percent of all data as our local testing set, which includes 42 HGG patients and 15 LGG patients. Representative segmentation results are shown in **Figure 4A**. The green shows the edema, the red shows the tumor necrosis, and the yellow shows the enhancing tumor. In order to evaluate the preliminary experimental results, we calculate the average Dice scores, sensitivity, and specificity for whole tumor, tumor core, and enhancing tumor, respectively. The results are shown in **Table 3**. The segmentation of whole tumor achieves best result with average Dice score of 0.8505.

Segmentation Results on MICCAI BraTS 2018 Validation Set of 66 Subjects

The segmentation results on BraTS 2018 online validation set achieve average Dice scores of 0.9048, 0.8364, and 0.7768 for whole tumor, tumor core, and enhancing tumor, respectively. That performance is slightly better than that in local testing set, while the whole tumor still has best result and enhancing tumor is the most challenging one. The details are shown in **Table 4**. For the ablation experiments, the distribution of Dice scores for whole tumor, tumor core and enhancing tumor are shown in **Figures 5A–C**, respectively. Generally, the average Dice scores for whole tumor, tumor core and enhancing tumor increase when ensembling more models to our cascaded V-Nets architecture. The difference of Dice scores for whole tumor between the baseline V-Nets architecture and our proposed architecture reaches significance as $p = 0.011$. Other model combination methods show the same trend although not get through Bonferroni correction.

Segmentation Results on MICCAI BraTS 2018 Testing Set of 191 Subjects

The segmentation results on BraTS 2018 online testing set achieve average Dice scores of 0.8761, 0.7953, and 0.7364 for whole tumor, tumor core and enhancing tumor, respectively. Compared with the Dice scores on MICCAI BraTS 2018 validation set, the numbers are slightly dropped. The details are shown in **Table 5**. The prediction of patient OS on BraTS 2018 testing set achieve accuracy of 0.519 and mean square error (MSE) of 367240. The details are shown in **Table 6**. The BraTS 2018 ranking of all participating teams in the testing data for both tasks has been summarized in Bakas et al. (2018), where our team listed as "LADYHR" and ranked 18 out of 61 in the segmentation task and 7 out of 26 in the prediction task.

Segmentation Results on Clinical Testing Sets of 56 Subjects

Representative segmentation results on two local hospital testing sets are shown in **Figures 4B,C**. The average Dice scores for whole tumor, tumor core and enhancing tumor in two data sets are calculated, respectively. The details are shown in **Table 7**. Overall, the images from China-Japan Union Hospital of Jilin University which are acquired using 2D MRI sequences achieve better segmentation results with Dice scores of 0.8635, 0.8036, and 0.7217 for whole tumor, tumor core, and enhancing tumor, respectively. On the other hand, the images from Affiliated Drum Tower Hospital of Nanjing University Medical School which are acquired using 3D MRI sequences achieve poor Dice score of 0.6786 for tumor core.

DISCUSSION

In this paper, we propose a cascaded V-Nets framework to segment brain tumor. The cascaded framework breaks down a difficult segmentation task into two easier subtasks including segmenting whole tumor from background and segmenting tumor substructures from whole tumor. Different from other methods, our method takes full account of the effect of preprocessing on the segmentation results, and use a customized preprocessing approach to process the data and train multiple models. The cascaded V-Nets are trained only using provided data, data augmentation and a focal loss formulation. We achieve state-of-the-art results on BraTS 2018 validation set. Specifically, the experimental results on BraTS 2018 online validation set achieve average Dice scores of 0.9048, 0.8364, and 0.7768 for whole tumor, tumor core and enhancing tumor, respectively. The corresponding values for BraTS 2018 online testing set are 0.8761, 0.7953, and 0.7364, respectively.

Generally, all the three average Dice scores degenerate in testing set compared with validation set. The reason may be that the sample size of testing set is much larger than that of validation set, and includes more anatomical variances. For clinical testing sets, we achieve 2% higher average Dice scores in images acquired using 2D MRI sequences than images acquired using 3D MRI sequences. The reason may be that the public dataset provided by the organizers of MICCAI BraTS 2018 includes more images acquired using 2D MRI sequences than images acquired using 3D MRI sequences. The trained model thus favors more 2D testing data than that of 3D. However, given that 2D MRI sequences are widely adopted in clinical practice for shorter acquisition

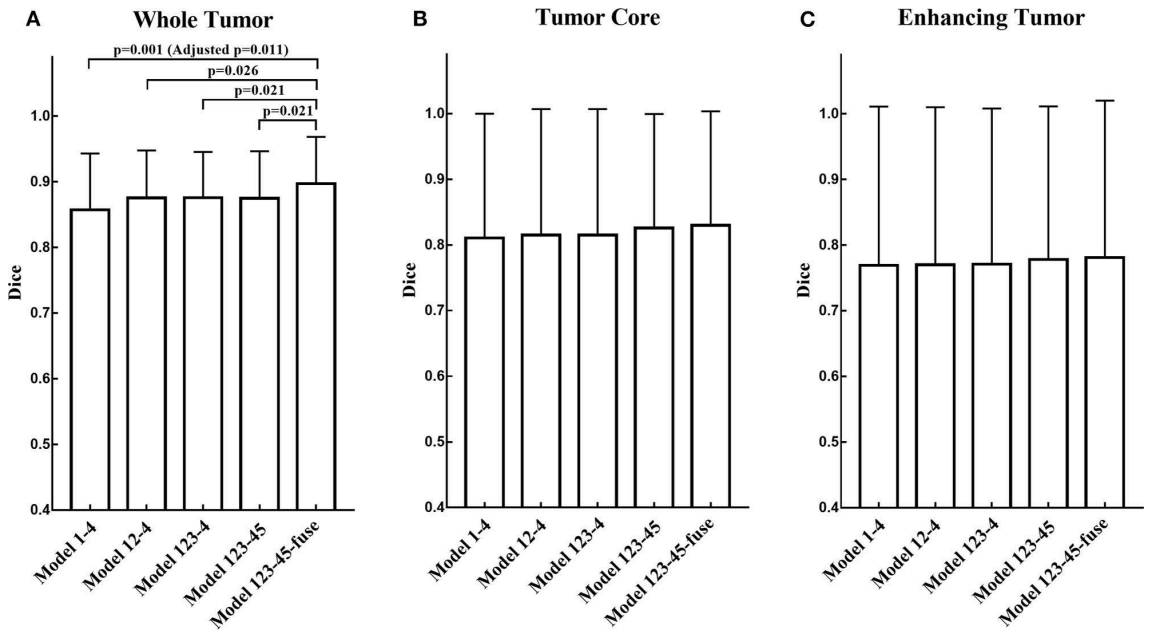


FIGURE 5 | The distribution of Dice scores for whole tumor, tumor core and enhancing tumor in ablation experiments. **(A)** The bar plot of Dice scores for whole tumor. The difference between the baseline V-Nets architecture and our proposed architecture reaches significance as $p = 0.011$. **(B)** The bar plot of Dice scores for tumor core. **(C)** The bar plot of Dice scores for enhancing tumor (The height of the bar indicates the mean Dice scores, and the error bars indicate the standard deviation).

TABLE 5 | Dice and Hausdorff95 measurements of the proposed method on BraTS 2018 testing set.

	Whole tumor	Tumor core	Enhancing tumor
Dice mean ± SD	0.8761 ± 0.1247	0.7953 ± 0.2543	0.7364 ± 0.2592
Hausdorff95 mean ± SD (mm)	7.0514 ± 11.5935	6.7262 ± 11.8852	3.9217 ± 6.1934

TABLE 6 | The prediction of patient overall survival on BraTS 2018 testing set.

	Scores
Accuracy	0.519
Mean squared error (MSE)	367239.974
Median square error (MedianSE)	38416
Standard deviation square error	945593.877
SpearmanR	0.168

TABLE 7 | Dice measurements of the proposed method on clinical testing set.

	China-Japan Union Hospital	Nanjing Drum Tower Hospital
# of subjects	28	28
Image resolution (mm^3)	0.6 × 0.6 × 6	0.67 × 0.67 × 0.67
WT Dice mean ± SD	0.8635 ± 0.0838	0.8692 ± 0.1307
TC Dice mean ± SD	0.8036 ± 0.1476	0.6786 ± 0.3093
ET Dice mean ± SD	0.7217 ± 0.1968	0.7054 ± 0.3557

time, the generated model may be more practical and meaningful. Therefore, for sites using major 3D images, the training set could include more 3D data and a specific 3D model could be trained.

There are several benefits of using a cascaded framework. First, the cascaded framework breaks down a difficult segmentation task into two easier subtasks. Therefore, a simple network V-Net can have excellent performance. In fact, in our experiment, V-Net does have better performance when segment the tumor substructures step by step than segment background and all the three tumor substructures together. Second, the segmentation results of V-Net 1 helps to reduce the receptive field from whole brain to only whole tumor. Thus, some false positive results can be avoided.

In addition to cascaded framework, ensemble strategy contributes to the segmentation performance. In our cascaded V-Nets framework, V-Net 1 includes models 1–3 and V-Net 2 includes models 4–5. Every model uses the same network structure V-Net. However, the training data is preprocessed with different pipelines mentioned in part of 2.1. According to our experimental experience, the Dice scores will greatly decrease due to the false positive results. While we did try several ways to change the preprocessing procedures for the training data, or change the model used in the segmentation task, the false positive results always appear. Interestingly, the false positive results appear in different areas in terms of different models. Therefore, ensemble strategy works by averaging probability maps obtained from different models. The results of the ablation experiments also confirm the proposed ensemble strategy works.

Moreover, we find three interesting points in the experiment. Firstly, for multimodal MR images, the combination of data

preprocessing procedures is important. In other words, different MRI modalities should be preprocessed independently. For example, in our first preprocessing pipeline, bias field correction only applied to T1 and T1Gd images. The reason is that the histogram matching approach may remove the high intensity information of tumor structure that has negative impact to the segmentation task. Secondly, we use three kinds of preprocessing methods to process the training and validation data, and compared their segmentation results. As a result, there is almost no difference between preprocessing methods in the three average Dice scores for whole tumor, tumor core and enhancing tumor, respectively. However, after the ensemble of the multiple models, the three average Dice scores all rose at least 2 percent. This suggests that data preprocessing methods is not the most important factor for the segmentation performance, while different data preprocessing methods are complementary and their combination can boost segmentation performance. Thirdly, the post-processing method is also important that it could affect the average Dices scores largely. If the threshold is too big, some of small clusters will be discarded improperly. If the threshold is too small, some false positive results will be retained. In order to have a better performance, we test a range of thresholds and choose the most suitable two thresholds as the upper and the lower bounds. For the components between upper and lower bounds, their average segmentation probabilities are calculated as a second criterion. Of course, these thresholds may not be suitable for all cases.

CONCLUSIONS

In conclusion, we propose a cascaded V-Nets framework to segment brain tumor into three substructures of brain tumor and background. The experimental results on BraTS 2018 online validation set achieve average Dice scores of 0.9048, 0.8364, and 0.7768 for whole tumor, tumor core and enhancing tumor, respectively. The corresponding values for BraTS 2018 online testing set are 0.8761, 0.7953, and 0.7364, respectively. The corresponding values for clinical testing set are 0.8635, 0.8036, and 0.7217, respectively. For clinical data set, images acquired using 2D MRI sequences achieve higher average Dice scores than images acquired using 3D MRI sequences, demonstrates that the proposed method is practical and meaningful in clinical practice. The state-of-the-art results demonstrate that V-Net is a promising network for medical imaging segmentation tasks, and the cascaded framework and ensemble strategy are efficient for boosting the segmentation performance.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was proved by the Institutional Review Board (IRB) of the Affiliated Drum Tower Hospital of Nanjing University Medical School and the China-Japan Union Hospital of Jilin University. Written informed consents were obtained from all subjects before participating to the study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported in part by the National Key Research and Development Program of China (2018YFC0116400) and the National Natural Science Foundation of China (81720108022, BZ). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Bakas, S, Reyes, M, Jakab, A, Bauer, A., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation progression assessment and overall survival prediction in the BRATS challenge. *arXiv 1811:02629*. doi: 10.17863/CAM.38755

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. *Cancer Imaging Arch.* 286. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017c). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Arch.* doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing The cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Cui, S., Mao, L., Jiang, J., Liu, C., and Xiong, S. (2018). Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network. *J. Healthc. Eng.* 2018:4940593. doi: 10.1155/2018/4940593

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago), 1026–1034. doi: 10.1109/ICCV.2015.123

Hua, R., Huo, Q., Gao, Y., Sun, Y., and Shi, F. (2019). "Multimodal brain tumor segmentation using cascaded V-Nets," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 49–60. doi: 10.1007/978-3-030-11726-9_5

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 287–297. doi: 10.1007/978-3-319-75238-9_25

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2019). "No New-Net," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 234–244. doi: 10.1007/978-3-030-11726-9_21

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2018). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI Brainlesion*

Workshop (Cham: Springer), 450–462. doi: 10.1007/978-3-319-75238-9_38

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv:1412.6980*.

Le Folgoc, L., Nori, A. V., Ancha, S., and Criminisi, A. (2016). "Lifted auto-context forests for brain tumour 335 segmentation," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 171–183. doi: 10.1007/978-3-319-55524-9_17

Lin, T., Goyal, P., Girshick, R., He, K., and Dollar, P. (2018). "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy), 2999–3007. doi: 10.1109/ICCV.2017.324

Mamelak, A. N., and Jacoby, D. B. (2007). Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (TM-601). *Expert Opin. Drug Deliv.* 4, 175–186. doi: 10.1517/17425247.4.2.175

Menze, B. H, Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Milletari, F., Navab, N., and Ahmadi, S. (2016). "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in Fourth *International Conference on 3D Vision (3DV)* (Stanford, CA), 565–571. doi: 10.1109/3DV.2016.79

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Wang, G. T., Li, W. Q., Ourselin, S., and Vercauteren, T. (2018). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 178–190. doi: 10.1007/978-3-319-75238-9_16

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., et al. (2012). Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. *Med. Image Comput. Comput. Assist. Interv.* 15, 369–376. doi: 10.1007/978-3-642-33454-2_46

Conflict of Interest: RH, QH, YG, and FS were employed by the company of Shanghai United Imaging Intelligence, Co., Ltd., Shanghai, China.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 18 February 2020
doi: 10.3389/fncom.2020.00010

A Novel Approach for Fully Automatic Intra-Tumor Segmentation With 3D U-Net Architecture for Gliomas

*Ujjwal Baid[1], Sanjay Talbar[1], Swapnil Rane[2], Sudeep Gupta[3], Meenakshi H. Thakur[4], Aliasgar Moiyadi[5], Nilesh Sable[4], Mayuresh Akolkar[4] and Abhishek Mahajan[4]**

[1] Department of Electronics and Telecommunication Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India, [2] Department of Pathology, Tata Memorial Centre, Tata Memorial Hospital, Mumbai, India, [3] Department of Medical Oncology, Tata Memorial Centre, Tata Memorial Hospital, Mumbai, India, [4] Department of Radiodiagnosis and Imaging, Tata Memorial Centre, Tata Memorial Hospital, Mumbai, India, [5] Department of Neurosurgery Services, Tata Memorial Centre, Tata Memorial Hospital, Mumbai, India

Purpose: Gliomas are the most common primary brain malignancies, with varying degrees of aggressiveness and prognosis. Understanding of tumor biology and intra-tumor heterogeneity is necessary for planning personalized therapy and predicting response to therapy. Accurate tumoral and intra-tumoral segmentation on MRI is the first step toward understanding the tumor biology through computational methods. The purpose of this study was to design a segmentation algorithm and evaluate its performance on pre-treatment brain MRIs obtained from patients with gliomas.

Materials and Methods: In this study, we have designed a novel 3D U-Net architecture that segments various radiologically identifiable sub-regions like edema, enhancing tumor, and necrosis. Weighted patch extraction scheme from the tumor border regions is proposed to address the problem of class imbalance between tumor and non-tumorous patches. The architecture consists of a contracting path to capture context and the symmetric expanding path that enables precise localization. The Deep Convolutional Neural Network (DCNN) based architecture is trained on 285 patients, validated on 66 patients and tested on 191 patients with Glioma from Brain Tumor Segmentation (BraTS) 2018 challenge dataset. Three dimensional patches are extracted from multi-channel BraTS training dataset to train 3D U-Net architecture. The efficacy of the proposed approach is also tested on an independent dataset of 40 patients with High Grade Glioma from our tertiary cancer center. Segmentation results are assessed in terms of Dice Score, Sensitivity, Specificity, and Hausdorff 95 distance (ITCN intra-tumoral classification network).

Result: Our proposed architecture achieved Dice scores of 0.88, 0.83, and 0.75 for the whole tumor, tumor core and enhancing tumor, respectively, on BraTS validation dataset and 0.85, 0.77, 0.67 on test dataset. The results were similar on the independent patients' dataset from our hospital, achieving Dice scores of 0.92, 0.90, and 0.81 for the whole tumor, tumor core and enhancing tumor, respectively.

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Madhura Ingalhalikar,
Symbiosis International University, India
Ashirbani Saha,
St. Michael's Hospital, Canada

***Correspondence:**
Abhishek Mahajan
drabhishek.mahajan@yahoo.in

Received: *31 August 2019*
Accepted: *27 January 2020*
Published: *18 February 2020*

Citation:
Baid U, Talbar S, Rane S, Gupta S, Thakur MH, Moiyadi A, Sable N, Akolkar M and Mahajan A (2020) A Novel Approach for Fully Automatic Intra-Tumor Segmentation With 3D U-Net Architecture for Gliomas. Front. Comput. Neurosci. 14:10. doi: 10.3389/fncom.2020.00010

Conclusion: The results of this study show the potential of patch-based 3D U-Net for the accurate intra-tumor segmentation. From experiments, it is observed that the weighted patch-based segmentation approach gives comparable performance with the pixel-based approach when there is a thin boundary between tumor subparts.

Keywords: glioma, intra-tumor segmentation, convolutional neural network, deep learning, 3D U-Net

INTRODUCTION

According to the Central Brain Tumor Registry of the United States (CBTRUS), 86,970 new cases of primary malignant and non-malignant brain tumors are expected to be diagnosed in the United States in 2019[1]. An estimated 16,830 deaths attributed to primary malignant brain tumors in the US in 2018. Gliomas are the most frequent primary brain tumors in adults and account for 70% of adult malignant primary brain tumors. Glioma arises from glial cells and infiltrates the surrounding tissues such as white matter fiber tracts with very rapid growth (Menze et al., 2015). Patients diagnosed with Glioblastoma tumors have an average survival time of 14 months (Louis et al., 2007).

Accurate segmentation of brain tumor tissues from Brain MR images is of profound importance in many clinical applications such as surgical planning and image-guided interventions (Mahajan et al., 2015). Manual tracing and detection of organs and tumor structure from medical images is considered as one of the preliminary steps in disease diagnosis, treatment planning, and monitoring tumor growth with follow-up evaluation (Udupa and Saha, 2003). In a clinical setup, this time-consuming process is carried out by radiologists, however, this approach becomes impractical when the number of patients increases. This presents an unmet need for automated segmentation methods (He et al., 2019; Vaidya et al., 2019).

In order to diagnose abnormality in brain tissues, various radio imaging techniques like Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Positron Emission Tomography (PET) are used. Over the last few decades, because of the better soft-tissue contrast, MRI is widely used to assess the brain tissues in clinical practices. Unlike X-rays or CT scans the intensity signature is variable in MRI due to various acquisition protocols. The same tumor cells follow different intensity distribution when acquired with different scanners with varying field strength, voxel resolution, and field of view. More accurate composite marking of the tumor regions can be achieved with four distinct MR sequences like T1, T2, T1 post-contrast (T1ce), and Fluid Attenuated Inversion Recovery (FLAIR). Intra-tumor parts for these four MR sequences with varying intensity can be visualized in **Figure 1**.

Different heterogeneous intra-tumor regions like edema, active tumor, and necrotic regions are present in Glial brain tumors. Intra-tumor segmentation in the brain has been challenging task because of its several characteristics such as non-rigid and complex appearance, variation in size, and position of tumor from patient to patient. Poor delineation of the intra-tumor parts in multi-modal MRI data as well as similar textural properties of the pathology with healthy tissues make the segmentation task more prone to error. It has been observed that even expert raters show significant variations in case of poor intensity gradients between tumor and rest of the healthy brain tissues. Though several algorithms have been proposed over the decades to address this task, most have shortcomings limiting their utility in routine clinical practice.

The aim of this study is to design a fully automated brain tumor segmentation algorithm which will accurately segment the tumors and act as an assistive tool for radiologists for exact tumor quantification. We have proposed a fully automatic brain tumor segmentation with 3D U-Net architecture based on Deep Convolutional Neural Networks. An efficient weighted patch extraction method along with a unique number of feature maps at each level of 3D U-Net is proposed for accurate intra-tumor segmentation.

We briefly review conventional and recent methods for brain tumor segmentation algorithms available in the literature. Further, BraTS challenge database along with local dataset from our hospital and proposed methodology for tumor segmentation is described. This is followed by experimental results, quantitative as well as a qualitative evaluation of the results and comparison with other methods. Finally, we conclude the manuscript with future directions.

LITERATURE REVIEW

As mentioned by Menze et al. there is a linear increase in the tumor imaging literature over the past 30 years and over 25% of the publications aimed at "automated" tumor segmentation. Segmentation of the glial tumors is the primary focus in most of the existing methods and very few methods targeted for specific glioma subtype or meningioma (Bauer et al., 2013). The brain tumor segmentation methods are broadly classified into two categories based on generative probabilistic based models and discriminative approaches. Generative probabilistic based approaches detect abnormal regions by comparing it with explicit models of anatomy and outlier detection. On the other hand, discriminative models learn from feature-based differences between normal tissues and tumor tissues.

Generative models aim at finding the outliers between *a-priori* model of a healthy brain (atlas) and the abnormal regions. This uses the prior information of tumor appearance and spatial distribution of the brain tissues and these methods exhibit good generalization to an unseen database (Prastawa et al., 2004). Cordier et al. (2016) proposed a fully automatic patch-based approach for Glioma segmentation with the multi-atlas

[1] Available: http://www.cbtrus.org/www.cbtrus.org/factsheet/factsheet.html

voting technique with less prior learning to avoid overfitting. The major drawback of these approaches is that it relies heavily on domain-specific prior knowledge and accurate multi-modal image registration. Because of the presence of large abnormalities and resection cavities in the brain, the multimodal registration miserably fails which lead to inaccurate segmentation in generative models.

Discriminative models directly learn from hand-designed features calculated on lesions and other brain tissues. This is carried out on large datasets to avoid the effect of imaging artifacts, intensity, and shape variations. In these approaches, various dense, and voxelwise features are extracted from MR images and fed into the classification algorithms like decision trees and support vector machines (Criminisi and Shotton, 2013). Demirhan et al. (2015) employed a method based on wavelets and Self-Organizing Maps (SOM) to segment intra-tumor parts along with healthy brain tissues. The drawback of these approaches is that, since the segmentation highly relies on the intensity, texture features etc. of the training data, segmentation is specific to the MRI images acquired with the same imaging protocol as of the training dataset.

Balafar et al. reviewed brain tumor segmentation methods and further classified them into four categories as Threshold-based, Region-based, Pixel classification based, and Model-based techniques with pros and cons over each other (Balafar et al., 2010). Many approaches to brain tumor segmentation have been implemented over decades but there is no winning theory.

Recent methods based on Deep Convolutional Neural Networks have outperformed all traditional machine learning methods in various domains like medical image segmentation, image classification, object detection, and tracking etc. (Smistad et al., 2015) and are currently considered to be art in biomedical image segmentation (Moeskops et al., 2016; Pereira et al., 2016; Havaei et al., 2017). The computational power of GPUs has enabled researchers to design deep neural network models with convolutional layers which are computationally expensive (Eklund et al., 2013; Eminaga et al., 2018; Lee et al., 2018; Leyh-Bannurah et al., 2018).

Pereira et al. (2016) proposed an automatic segmentation method using Convolutional Neural Networks by exploring smal 3 × 3 kernels. 2D patches were extracted from four MR channels of size 33×33 for training the network. Ronneberger et al. (2015) segmented the neuronal structures in electron microscopic stacks with 2D U-Net architecture trained on transmitted light microscopy images with augmentation of the training data by geometrical image transformations. Kamnitsas et al. (2017) proposed dual pathway architecture with dense training scheme to incorporate both local and larger contextual information. The architecture processed the input images at multiple scales simultaneously. False positives in the segmentation maps were minimized using Conditional Random Forests (CRF).

Inspired from the above literature, we developed a novel Deep Convolutional Neural Network-based 3D U-Net model with a unique number of feature maps. Various heterogeneous histologic sub-regions like peritumoral edema, enhancing tumor, and necrosis were accurately segmented in spite of thin and/or fuzzy boundaries between intra-tumor parts with this proposed architecture.

PATIENTS AND METHOD

We focused our experimental analysis on MICCAI (Medical Image Computing and Computer-Assisted Intervention) Brain Tumor Segmentation (BraTS) 2018 challenge (Bakas et al., 2019). BraTS dataset consisted of multi-institutional routine clinically acquired pre-operative multimodal MRI scans of High Grade Glioma i.e., Glioblastoma (GBM/HGG) and Lower Grade Glioma (LGG), with a pathologically confirmed diagnosis. In the challenge, MR data of 285 patients for training, 66 for validation and 191 patients were provided in the test dataset. The MR data was acquired with different imaging clinical protocols and various MR scanners with 19 distinct institutions (Bakas et al., 2017a,b). Each patient data was provided with FLAIR, T1, T2, T1 post-contrast MR volume of size 240×240×155 which were resampled to 1 mm × 1 mm × 1 mm resolution. Segmentation labels as edema, enhancing tumor, and necrosis were annotated for all patients by one to four radiologists as shown in **Figure 1**. These segmented labels were also verified by expert neuro-radiologists. The main task of BraTS 2018 challenge was to auto-segment the tumor into its three constituent regions viz.

1. Enhancing tumor region (ET)
2. Tumor Core (TC) which entails the ET, necrotic (fluid-filled) and the non-enhancing (solid) parts
3. Whole tumor (WT) which includes all intra-tumor parts along with Edema.

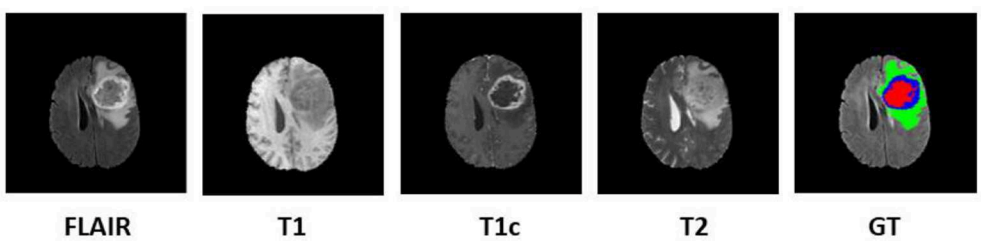


FIGURE 1 | Multi-modal data with four channels provided in BraTS 2018 challenge dataset along with Ground Truth (GT). Sub tumor parts are represented as—Edema: Green, Enhancing tumor: Blue, Necrosis: Red.

Apart from BraTS 2018 dataset, the proposed method was also tested on 40 pre-treatment multimodal MRI patient datasets of Glioblastoma (GBM) from our hospital. MR data of four channels as FLAIR, T1, T2, and T1 post contrast was collected for the study. The acquisition protocol is provided in **Supplementary Material**. The local dataset was explicitly used for the purpose of testing only. This dataset was also skull-stripped and resampled to 1 mm × 1 mm × 1 mm resolution. This dataset was annotated by the expert radiologists from our hospital with the same protocol which was defined to annotate BraTS challenge dataset (Menze et al., 2015).

Pre-processing

The input data for the segmentation algorithm were skull stripped, normalized, and co-registered to an anatomical template (Smith, 2002). In order to normalize the signal intensities between the BraTS and our hospital datasets, bias field correction was performed with N4ITK tool (Tustison et al., 2010). Further, MR data of each channel was normalized by subtracting the channel mean and dividing by the variance i.e., zero mean and unit variance.

Patch Extraction

Tumor sub-region distribution in BraTS training data was highly imbalanced. Further, 98% pixels of the dataset belonged to either healthy brain tissues or background and hence the model was prone to overfit on non-tumor tissues only. The problem was exaggerated when the prediction was made based on center pixel class of the patch. Hence, precise patch selection from the input data for training is of extreme importance. To overcome this problem, we adopted a novel 3D patch-based approach for training with weighted sampling. Zhou et al. (2019) reviewed 2D and 3D patch extraction methods along with several types of loss functions. The main approaches included resampling the data space as: under-sampling the negative class or up-sampling the negative class and SMOTE (Synthetic Minority Over-sampling Technique) generating synthetic samples. The methods discussed includes patch extraction 50% probability being cantered either on the lesion or healthy voxel. Also, all training patches centered on a lesion voxel. AlBadawy et al. (2018) discussed impact of cross institutional training and testing for segmentation of brain tumors. In this study patches of 33*33 were extracted on T1, T1ce, and FLAIR modality. Our proposed approach differs with this approach in terms of dimension of patch size as 64*64*64. It is well-known fact that T2 modality is widely used to distinguish tumor core boundary with rest of the tumor and hence we included T2 channel as well along with the other three MR channels to incorporate more information during training.

In our proposed approach, 3D patches were extracted from all the four modalities so that the network can be trained on a distinct intensity signature of intra-tumor tissues in each modality. For this, we considered the equidistant seed points in X, Y, and Z directions of the MR data as shown in **Figure 2**. A 3D patch of size 64 × 64 × 64 voxels was considered around each seed point. In the next step, potential patches which had brain area more than 60% of the total patch were only considered for the training to minimize the chances of overfitting of the model to the background pixels. It was observed that the model was misclassifying the pixels on tumor boundary to healthy brain tissues. A similar problem occurred when tumors were present on the boundary of the brain, with pixels being classified to background. To address this, some patches were explicitly extracted on the boundary of the tumor with weighted sampling as shown in **Figure 3**. The boundary locations of the WT is considered as the tumor boundary to extract the additional patches. This is done with find_boundaries() function available in segmentation module in popular skimage library. Randomly 30% boundary locations are selected for these extra patch extractions. Since, there is high class imbalance in tumor tissues and healthy tissues, this additional patch extraction does not impact on the performance of model like biased training or overfitting. These additional patches were added to the training patch dataset so that model could be trained in a better way to distinguish thin

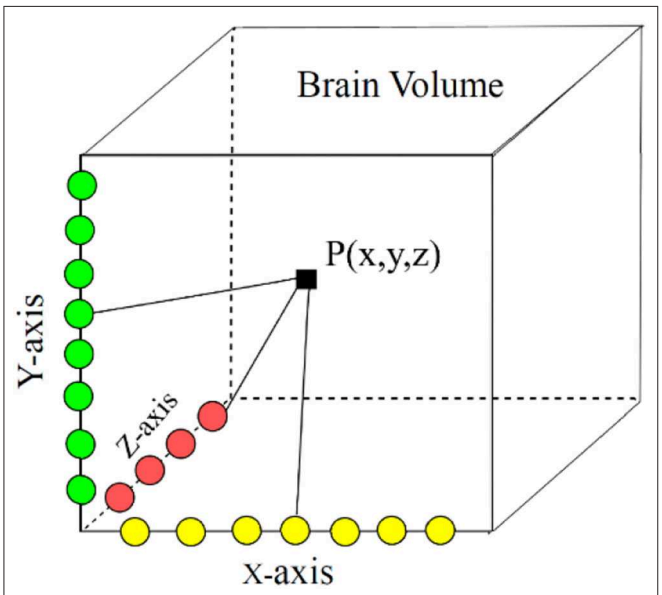


FIGURE 2 | Patch center localization by randomly selecting x, y, z coordinates in brain volume.

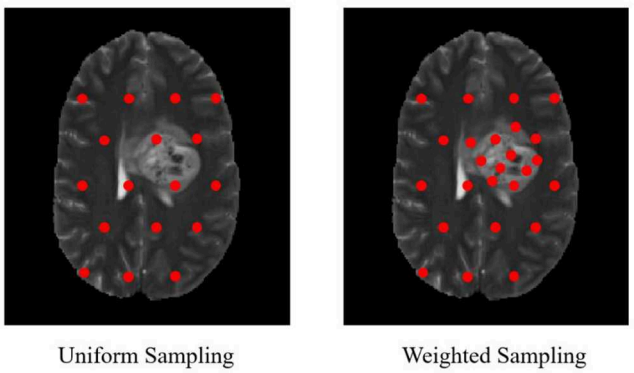


FIGURE 3 | Uniform sampling and weighted sampling for patch extraction.

boundaries of the tumor with the rest of the brain or background. This weighted patch extraction pipeline is fully automatic i.e., without any manual intervention. These 3D patches from all the four channels were concatenated together and given as input to the first layer of the model along with corresponding ground truth during training. During testing as well the non-overlapping patches of size 64 × 64 × 64 were extracted and final output volume is generated by concatenating all these predicted patches to get single 240 × 240 × 155 volume.

Proposed 3D U-Net Architecture

Conventional U-Net architecture consists of a bunch of basic layers such as convolutional layers, down-sampling and upsampling layers etc. Several variants of the 2D and 3D U-Net architectures are available in the recent literature which mainly differ in respect to the choice of hyperparameters viz. depth of U-Net, number of feature maps, kernel size etc. Selection of these hyperparameters along with accurate region input is of utmost importance for accurate training of the model. The novelty of our proposed approach lies in the weighted patch extraction scheme from the edges of the tumor and designing the structure of 3D U-net with less number of levels and an increased number of filters at each level. Although several deeper U-Net architectures are proposed for segmentation task, we restricted our network to three levels. This reduced the number of trainable parameters but also avoided the bottleneck problem caused due to smaller patch size.

In proposed 3D U-Net architecture, from the first level to third level 48, 96, and 192 feature maps were present at each subsequent level in down-sampling and up-sampling layers as shown in **Figure 4**. The proposed architecture consisted of a contracting path to capture context and a symmetric expanding path that enables precise localization. At the first layer four 64 × 64 × 64 multichannel MR volume data was given as input for training along with the corresponding ground truth. The number of features maps increased in the subsequent layers to learn the deep tumor features. These were followed by ReLU activation function and the features were down-sampled in encoding layer. Similarly, in decoding layer after convolution layers and ReLU activation function, features maps were up-sampled by a factor of 2. Features maps from encoding layers were concatenated to the corresponding decoding layer in the architecture. In contrast to conventional U-Net, all the feature maps were zero-padded to keep the same output dimensions for all convolutional layers. Finally, four output maps were generated with 1 × 1 convolutional layer corresponding to non-tumor tissue, edema, necrosis, and enhancing tumor. Each voxel of these four output maps corresponds to the probability of each voxel belonging to the particular class. The final prediction was generated by selecting the label with maximum probability from these four label maps. At the output layer, the segmentation map predicted by the model was compared with the corresponding ground truth and the error was backpropagated in the intermediate 3D U-Net layers.

In our implementation, the learning rate (α) was initialized to 0.001 and remained unchanged till 60 epochs. Since, after 60 epochs the Dice loss stopped improving, we decreased it linearly by a factor of 10^{-1} which avoided convergence of the model to local minima. The model is trained for 100 epochs since beyond that there was no significant improvement in the Dice loss and hence the training was terminated. Dropout with ratio 0.25 was added during training to avoid overfitting. The architecture was trained with a batch size of 8. Further, for better optimization a momentum strategy was included in the implementation. This used a temporally averaged gradient to damp the optimization velocity.

Post-processing

False positives in the segmentation output within the brain region were minimized with 3D Connected Component Analysis with the largest connected component being retained in each predicted volume. Similarly, false positives from the background were eliminated using a binary brain mask generated from brain volume and overlaid on the segmentation output with a logical

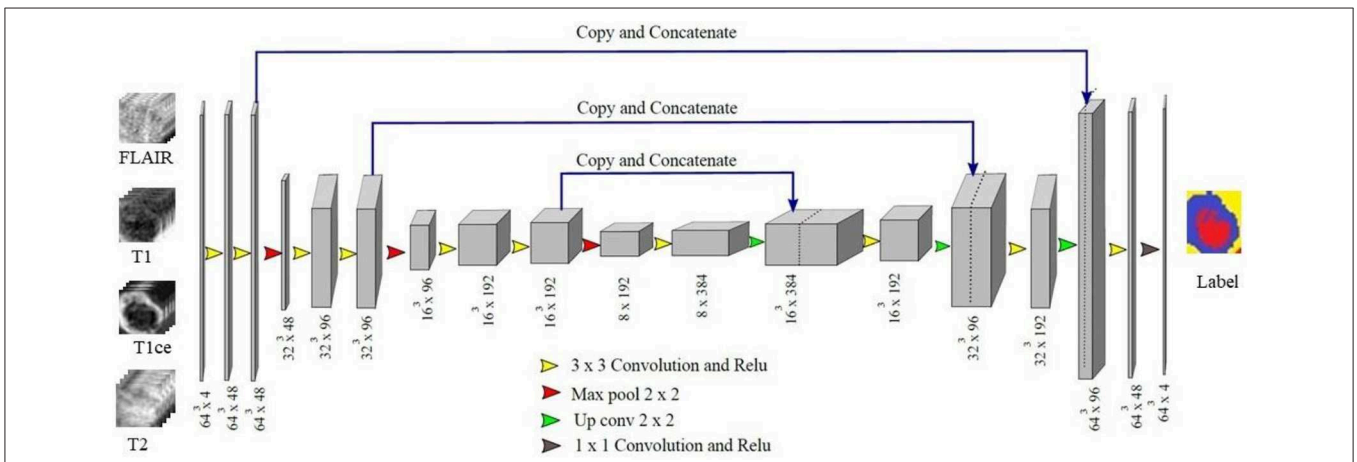


FIGURE 4 | Proposed 3D U-Net Architecture. Voxels from all four MR channels were given input to the first layer of the model. The predicted labels were compared with the Ground truth to calculate Dice loss.

AND operation. This improved the accuracy of the segmentation significantly for tumors present on the boundaries of the brain. There are some limitations of 3D connected component analysis as post-processing method where bifocal tumors are present that too in distinct brain lobes.

Implementation Details

The proposed architecture was implemented using Tensorflow library which supported the use of GPUs (Agarwal et al., 2015). GPU implementation greatly accelerated the implementation of deep learning algorithms. The approximate time to train the model was 48 h on 16 GB NVIDIA P100 GPU using cuDNN v5.0 and CUDA 8.0 with 128 GB RAM. The prediction on validation data took <60 s for a single patient with four MR channels data, each of dimension 240 × 240 × 155.

RESULTS AND DISCUSSION

The quantitative evaluation of the proposed model was done on BraTS 2018 challenge dataset and also on an independent

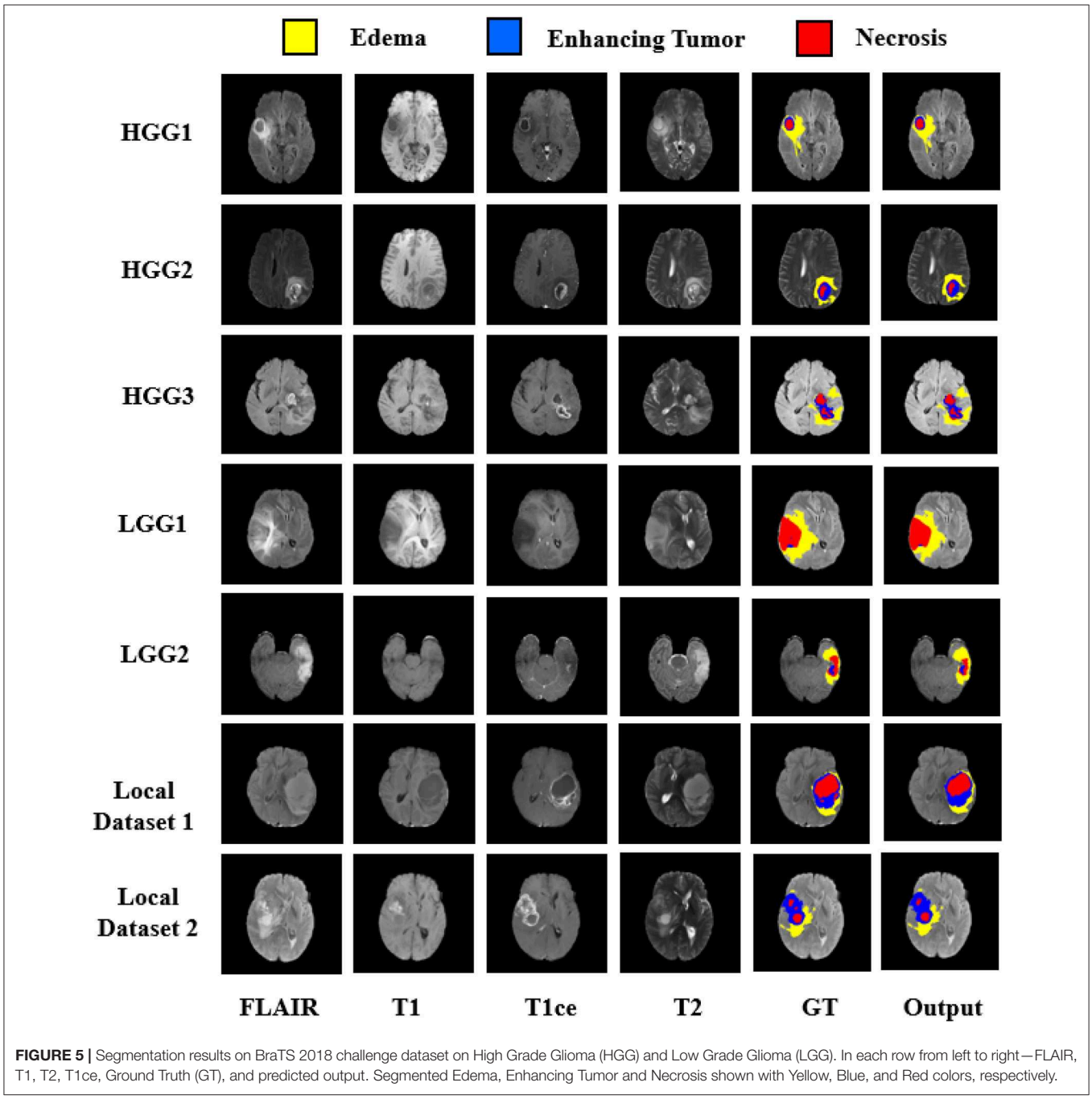


FIGURE 5 | Segmentation results on BraTS 2018 challenge dataset on High Grade Glioma (HGG) and Low Grade Glioma (LGG). In each row from left to right—FLAIR, T1, T2, T1ce, Ground Truth (GT), and predicted output. Segmented Edema, Enhancing Tumor and Necrosis shown with Yellow, Blue, and Red colors, respectively.

dataset of GBMs from our hospital. The BraTS dataset comprised of three data sub-sets, viz. training, validation, and test dataset. No ground truths were provided for validation and test dataset. The representative results on BraTS challenge dataset are shown in **Figure 5** with High Grade Glioma (HGG) and Low Grade Glioma (LGG). Edema, Enhancing Tumor, and Tumor Core segmented by our approach are shown with Yellow, Blue, and Red colors, respectively.

Quantitative Performance Evaluation

Performance evaluation was done based on Dice Score, Sensitivity, Specificity, and Hausdorff 95 distance. These evaluation matrices are measures of voxel-wise overlap of the segmented regions (CBICA Image Processing Portal[2]; Taha and Hanbury, 2015). The Dice score normalizes the number of true positives to the average size of the two segmented areas. It is identical to the F score (the harmonic mean of the precision-recall curve) and can be transformed monotonously to the Jaccard score. For the tumor regions Dice Score, Sensitivity (True positive rate), and Specificity (True negative rate) were computed as shown in Equations (1)–(3).

$$Dice\,(P,T) = \frac{2 * |P_1 \cap T_1|}{(|P_1| + |T_1|)} \quad (1)$$

$$Sensitivity\,(P,T) = \frac{|P_1 \cap T_1|}{(|T_1|)} \quad (2)$$

$$Specificity\,(P,T) = \frac{|P_0 \cap T_0|}{(|T_0|)} \quad (3)$$

Where, P represents the model prediction and T represents the Ground Truth labels. T_1 and T_0 are the subset of voxels predicted as positive and negatives for tumor region and similar for P_0 and P_1 as shown in **Figure 6**. The Hausdorff 95 distance is the 95th quartile of the maximum overall surface distance between predicted surface and ground truth surface. Hausdorff 95 overcomes the problem of high sensitivity of the Hausdorff measure to small outlying sub-regions from both P1 and T1 (Taha and Hanbury, 2015). Specificity was also calculated and was noted to be >99% in all the cases. Mean, median, standard deviation, 25 quartile, and 75 quartile were also computed for all the patients in the dataset. The BraTS challenge organizers had provided online evaluation system for all the training, validation, and test cases from the BraTS dataset (CBICA Image Processing Portal). The evaluation metrics were calculated by us for the in-house cases (**Table 1** and **Figure 7**).

In BraTS 2018 training dataset, the mean Dice score for ET, WT, and TC was 0.80, 0.93, and 0.91, respectively. The model predicted ET, WT, and TC with Dice score of 0.75, 0.88, and 0.83 for validation dataset. Comparison of our approach with other methods participated in the BraTS challenge is given in **Table 2**.

[2]CBICA Image Processing Portal. Available: https://ipp.cbica.upenn.edu/

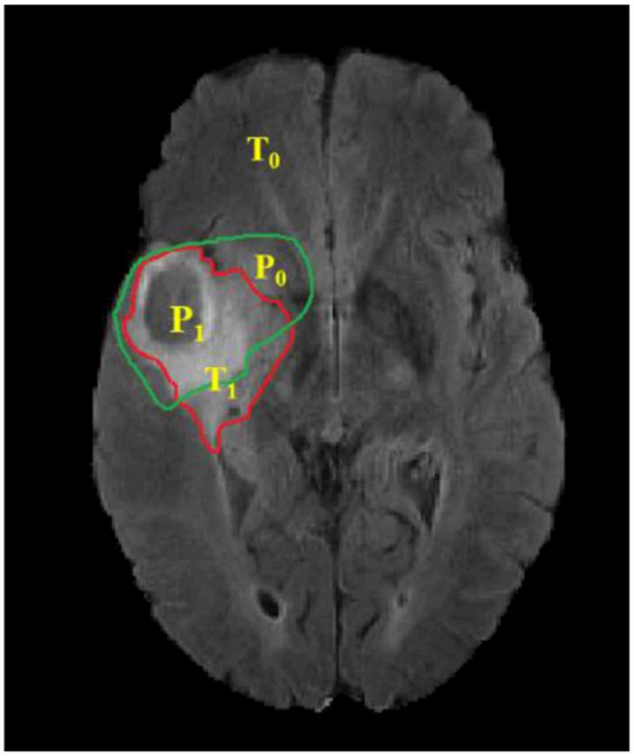


FIGURE 6 | Red contour: ground truth, green contour: predicted segmentation. Notation T is to denote ground truth and P to the predicted segmentation output.

From **Table 2**, it can be observed that our approach achieved better segmentation accuracy in terms of Dice Score over other methods available in the literature. We tested the proposed architecture on 40 patients from our hospital and achieved Dice Score 0.81, 0.92, and 0.90 for ET, WT, and TC, respectively.

The proposed approach outperformed over other U-net based deep learning approaches available in the literature as shown in **Table 2** for training and validation dataset. Since the performance of other methods on test dataset are not available publically and hence not included in the comparison. As different tumor parts appear with distinct intensities in FLAIR, T1, T2, and T1ce modalities, we extracted 3D patches from all the four modalities which resulted in better training for intra-tumor segmentation. Also, we resolved the problems resulting due to focusing on the center pixels of a patch as has been the norm in previous approaches (Pereira et al., 2016) which results in high misclassification due to severe class imbalance in the patches. We instead have merged the four segmentation label maps corresponding to enhancing tumor, necrosis, edema, and background predicted at the output layer, to generate a single segmentation map.

High class imbalance is also intrinsic to most imaging datasets. Around 98.88% pixels belonged to background/healthy class while an average of 0.64, 0.20, and 0.23% pixels belonged to Edema, enhancing tumor and necrosis, respectively. Training of the model with this class imbalance would result in overfitting to

TABLE 1 | Experimental results on BraTS 18 challenge training, testing, and validation dataset.

Datasets	Evaluation parameters	Dice			Hausdorff 95		
		ET	WT	TC	ET	WT	TC
BraTS18 Training (285 patients)	Mean	0.8202	0.9324	0.9198	7.0750	11.0278	11.0985
	SD	0.2746	0.1057	0.1327	21.2342	27.9139	29.4150
	Median	0.9062	0.9614	0.9565	1.0000	1.4142	1.0000
	25 Quartile	0.8422	0.9406	0.9303	1.0000	1.0000	1.0000
	75 Quartile	0.9422	0.9728	0.9687	1.4142	1.7320	2.0000
BraTS18 Validation (66 patients)	Mean	0.7480	0.8780	0.8267	7.2951	16.8157	11.2021
	SD	0.2659	0.1346	0.1828	15.7042	30.2509	20.2365
	Median	0.8527	0.9180	0.8985	2.2360	3.3131	4.3589
	25 Quartile	0.7325	0.8665	0.7771	1.4142	2.0000	2.0000
	75 Quartile	0.8853	0.9420	0.9444	3.9354	8.4183	9.4868
BraTS18 Testing (191 patients)	Mean	0.6677	0.8475	0.7688	9.0554	17.2184	14.5728
	SD	0.3120	0.1699	0.2786	19.8975	28.9190	26.1504
	Median	0.8013	0.9050	0.8946	2.2360	3.4641	3.3166
	25 Quartile	0.6557	0.8336	0.7519	1.4142	2.2360	2.0000
	75 Quartile	0.8657	0.9404	0.9328	3.6055	9.4604	8.4844
Our patient dataset (40 patients)	Mean	0.8134	0.9235	0.9012	6.0863	8.1789	9.8647

ET, enhancing tumor; WT, whole tumor; TC, tumor core; SD, standard deviation.

the healthy class leading to misclassification of necrotic pixels to healthy pixels. This problem was overcome by weighted sampling and augmenting the data for under-represented regions. Patches from the boundary region of the tumor were added explicitly for better training of the model with weighted patch extraction. All these steps increased the segmentation accuracy at the tumor boundaries.

In patch-based training approaches, larger patches require more max-pooling layers which minimize the localization accuracy. Contrarily, training with small patches allows the network to see only little context. Hence, a classifier output that takes into account the features from multiple layers is considered. This leads to better localization with the use of context. We experimented with various patch extraction size and schemes along with variations in encoding and decoding layers in terms of number and dimension of the Conv-filters. We finalized various hyperparameters like the number of Conv layers, feature maps, activation function, loss function, patch size, learning rate, etc. by extensive experimentation on validation dataset. We evaluated the performance of the model on online evaluation portal for validation dataset and the hyperparameters for which best validation Dice score is achieved are finalized. Three encoding and three decoding layers with 48, 96, and 192 feature maps with ReLU activation function is used in the model with training on patch size of $64 \times 64 \times 64$. The weights of the proposed model are updated according to Dice loss. Some notable variations and performance are provided in **Supplementary Table 2**.

Box plot for all the patients in BraTS training and validation dataset are shown in **Figure 7**. It can be observed that the median value is much higher than the mean value in terms of Dice Score. Theoretically, Dice Score ranges from minimum 0 to maximum 1. From the box plots, it can be observed that the Dice scores of two cases for enhancing tumor and tumor core segmentation results are very close to 0 and for the whole tumor is below 0.5 in a few cases. These regions failed to segment accurately because of the high deviation in characteristics in training and validation dataset. This problem can be overcome by increasing the training data with inter-patient variations.

Grading of Segmentation by Neuroradiologist

The segmentation results on the in-house testing dataset were further evaluated by an in-house expert radiologist (AM) on a scale of 0–5. Score 0 referred to the poor segmentation and 5 for the most accurate delineating of the tumor parts from the healthy tissues. The subjective score for almost all segmented images was found acceptable by the radiologists. However, in a few cases with large necrotic tumor cavity, the proposed algorithm failed to accurately segment the tumor parts. We further investigated the problem and found that such cases were not present in the BraTS challenge training dataset on which proposed architecture was trained and this can be addressed by increasing the training dataset with patients belonging such type of tumor parts. We achieved average 4.1 and median 4 score by the expert Neuroradiologist. The details are provided in **Supplementary Table 1**. Since, BraTS validation and dataset comprised of the scans from multiple institution with varying protocols the performance on them is comparatively poor. Also, it was observed that on online evaluation portal even if you predict a single pixel for the sub-tumor part which is not present in the patient scan, the Dice score for the corresponding case is zero which reduces the mean Dice score on complete dataset. MR

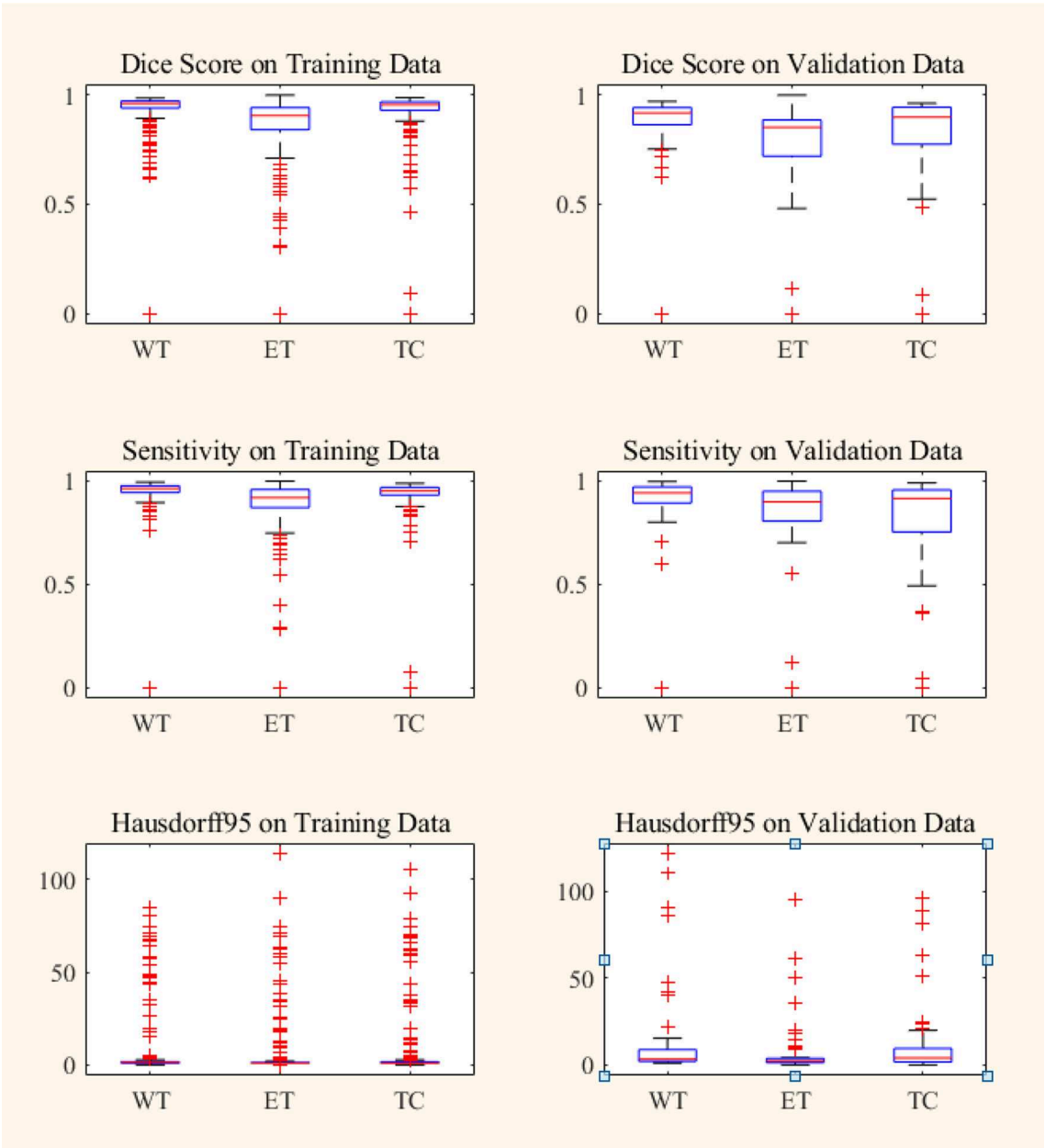


FIGURE 7 | Box plot of Dice score, Sensitivity and Hausdorff 95 distance on BraTS18 training and validation data. Red line within box plot is the median of the corresponding data. ET, enhancing tumor; WT, whole tumor; TC, tumor core.

data of all the patients from our in-house dataset was with all the tumor subparts and hence there were no cases for which the Dice score as zero.

CONCLUSION

In this paper, we presented fully automatic brain tumor segmentation with a novel 3D U-Net architecture based on Deep Convolutional Neural Networks. An efficient weighted patch extraction method along with a unique number of feature maps at each level of 3D U-Net is proposed for accurate intra-tumor segmentation. The performance of the proposed algorithm is evaluated on BraTS 2018 dataset as well as on the dataset from the local hospital. We considered different training schemes with variable patch sizes, data augmentation methods, activation functions, loss functions, and optimizers. Nowadays, adversarial networks are outperforming state of the art methods for semantic segmentation in several Computer Vision tasks. This can be further inverstigated to improve the segmentation in medical images. The work can also be extended for prediction of overall

TABLE 2 | Comparison of proposed architecture with other segmentation methods who participated in BraTS 2018 challenge.

BraTS18 datasets	References	Dice			Hausdorff 95		
		ET	WT	TC	ET	WT	TC
Validation	Cabezas et al., 2018	0.7403	0.8892	0.7200	5.3035	6.9563	11.9238
	Chen et al., 2018	0.7334	0.8878	0.8078	4.6426	5.50541	8.14015
	Fang and He, 2018	0.7200	0.8560	0.7260	5.7000	7.5000	9.5000
	Gates et al., 2018	0.6783	0.8055	0.6852	14.5229	14.4150	20.0174
	Hu et al., 2018	0.6100	0.8300	0.7300	41.4800	47.2300	41.1400
	Myronenko, 2019	0.8233	0.9100	0.8668	3.9257	4.5160	6.8545
	Isensee et al., 2018	0.8087	0.9126	0.8634	2.41	4.27	6.52
	Mehta and Tal, 2019	0.7880	0.9090	0.825	3.520	4.923	8.316
	Lefkovits et al., 2018	0.7190	0.8730	0.6890	7.3040	7.0680	12.6630
	Proposed 3D U-Net	0.7480	0.8780	0.8267	7.2951	12.9486	11.2021

ET, enhancing tumor; WT, whole tumor; TC, tumor core.

survival prediction of the patient with the radiomic features computed on the predicted tumor.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2018/data.html.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Multimodal Brain Tumor Segmentation Challenge 2018. As this study was carried out retrospectively on pre-existing data, written informed consent for participation was not required in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

UB, AMa, and SR conducted the experiment. All the authors contributed to writing the manuscript and are responsible.

FUNDING

This work was supported by Ministry of Electronics and Information Technology, Govt. of India under Visvesvaraya Ph.D. scheme with implementation reference number: PhD-MLA/4(67/2015-16).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00010/full#supplementary-material

REFERENCES

Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Available online at: https://arxiv.org/abs/1603.04467

AlBadawy, E. A., Saha, A., and Mazurowski, M. A. (2018). Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. *Med. Phys.* 45, 1150–1158. doi: 10.1002/mp.12752

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Can. Img. Arch.* doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., and Rempfler, M. (2019). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv Prepr. arXiv*1811.01328.

Balafar, M. A., Ramli, A. R., Saripan, M. I., and Mashohor, S. (2010). Review of brain MRI image segmentation methods. *Artif. Intell. Rev.* 33, 261–274. doi: 10.1007/s10462-010-9155-0

Bauer, S., Wiest, R., Nolte, L. P., and Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* 58, R97–R129. doi: 10.1088/0031-9155/58/13/R97

Cabezas, M., Valverde, S., Gonzalez-Villa, S., Clerigues, A., Salem, M., Kushibar, K., et al. (2018). "Survival prediction using ensemble tumor segmentation and transfer learning," in *Pre-conference Proceedings of the 7th Medical Image Computing and Computer-Assisted Interventions (MICCAI) BraTS Challenge 2018* (Granada), 54–62.

Chen, W., Liu, B., Peng, S., Sun, J., and Qiao, X. (2018). "S3D-UNet: separable 3D U-Net for brain tumor segmentation," in *Pre-conference Proceedings of the 7th Medical Image Computing and Computer-Assisted Interventions (MICCAI) BraTS Challenge 2018*, 91–99.

Cordier, N., Delingette, H., and Ayache, N. (2016). A patch-based approach for the segmentation of pathologies: application to glioma labelling. *IEEE Trans. Med. Imaging* 35, 1066–1076. doi: 10.1109/TMI.2015.2508150

Criminisi, A., and Shotton, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Publishing Company. doi: 10.1007/978-1-4471-4929-3

Demirhan, A., Toru, M., and Guler, I. (2015). Segmentation of tumor and edema along with healthy tissues of brain using wavelets and neural networks. *IEEE J. Biomed. Heal. Informatics* 19, 1451–1458. doi: 10.1109/JBHI.2014.2360515

Eklund, A., Dufort, P., Forsberg, D., and LaConte, S. M. (2013). Medical image processing on the GPU - Past, present and future. *Med. Image Anal.* 17, 1073–1094. doi: 10.1016/j.media.2013.05.008

Eminaga, O., Eminaga, N., Semjonow, A., and Breil, B. (2018). Diagnostic classification of cystoscopic images using deep convolutional neural networks. *JCO Clin. Cancer Inform.* 2, 1–8. doi: 10.1200/CCI.17.00126

Fang, L., and He, H. (2018). "Three pathways U-Net for brain tumor segmentation," in *Pre-conference proceedings of the 7th medical image computing and computer-assisted interventions (MICCAI) BraTS Challenge* 2018, 119–126.

Gates, F., Pauloski, J. G., and Schellingerhout, D., and Fuentes, D. (2018). "Glioma segmentation and a simple accurate model for overall survival prediction," in *Pre-conference Proceedings of the 7th Medical Image Computing and Computer-Assisted Interventions (MICCAI) BraTS Challenge 2018*, 144–152.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

He, Y., Zhang, H., and Wang, Y. (2019). *RawNet: Fast End-to-End Neural Vocoder*. Available online at: https://arxiv.org/abs/1904.05351.

Hu, Y., Liu, X., Wen, X., Niu, C., and Xia, Y. (2018). "Brain tumor segmentation on multimodal MRI using multi-level upsampling in decoder," in *Pre-conference Proceedings of the 7th Medical Image Computing and Computer-Assisted Interventions (MICCAI) BraTS Challenge 2018*, 196–204.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Granada), 234–244.

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Lee, J., An, J. Y., Choi, M. G., Park, S. H., Kim, S. T., Lee, J. H., et al. (2018). Deep learning–based survival analysis identified associations between molecular subtype and optimal adjuvant treatment of patients with gastric cancer. *JCO Clin. Cancer Inform.* 2, 1–14. doi: 10.1200/CCI.17.00065

Lefkovits, S., Szilágyi, L., and Lefkovits, L. (2018). "Cascade of random forest classifiers for brain tumor segmentation," in *Pre-conference Proceedings of the 7th Medical Image Computing and Computer-Assisted Interventions (MICCAI) BraTS Challenge 2018*, 280–289.

Leyh-Bannurah, S. R., Tian, Z., Karakiewicz, P. I., Wolffgang, U., Sauter, G., Fisch, M., et al. (2018). Deep learning for natural language processing in urology: state-of-the-art automated extraction of detailed pathologic prostate cancer data from narratively written electronic health records. *JCO Clin. Cancer Inform.* 2, 1–9. doi: 10.1200/CCI.18.00080

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., et al. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* 114, 97–109. doi: 10.1007/s00401-007-0243-4

Mahajan, A., Moiyadi, A. V., Jalali, R., and Sridhar, E. (2015). Radiogenomics of glioblastoma: a window into its imaging and molecular variability. *Cancer Imaging* 15(Suppl 1), 5–7. doi: 10.1186/1470-7330-15-S1-P14

Mehta, R., and Tal, A. (2019). *3D U-Net for Brain Tumour*. Granada: Springer International Publishing.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., and Isgum, I. (2016). Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans. Med. Imaging* 35, 1252–1261. doi: 10.1109/TMI.2016.2548501

Myronenko, A. (2019). *3D U-Net for Brain Tumour*. Granada: Springer International Publishing.

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* 35, 1240–1251. doi: 10.1109/TMI.2016.2538465

Prastawa, M., Bullitt, E., Ho, S., and Gerig, G. (2004). A brain tumor segmentation framework based on outlier detection. *Med. Image Anal.* 8, 275–283. doi: 10.1016/j.media.2004.06.007

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015* (Munich), 234–241.

Smistad, E., Falch, T. L., Bozorgi, M., Elster, A. C., and Lindseth, F. (2015). Medical image segmentation on GPUs - a comprehensive review. *Med. Image Anal.* 20, 1–18. doi: 10.1016/j.media.2014.10.012

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Taha, A. A., and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* 15:29. doi: 10.1186/s12880-015-0068-x

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Udupa, J. K., and Saha, P. K. (2003). Fuzzy connectedness and image segmentation. *Proc. IEEE* 91, 1649–1669. doi: 10.1109/JPROC.2003.817883

Vaidya, T., Agrawal, A., Mahajan, S., Thakur, M. H., and Mahajan, A. (2019). The continuing evolution of molecular functional imaging in clinical oncology: the road to precision medicine and radiogenomics (part I). *Mol. Diagnosis Ther.* 23, 27–51. doi: 10.1007/s40291-018-0367-3

Zhou, T., Ruan, S., and Canu, S. (2019). A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* 3–4:100004. doi: 10.1016/j.array.2019.100004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fnins.2020.00027

Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches

Tahsin Kurc[1†‡], Spyridon Bakas[2,3,4†‡], Xuhua Ren[5], Aditya Bagari[6], Alexandre Momeni[7], Yue Huang[8], Lichi Zhang[5], Ashish Kumar[6], Marc Thibault[7], Qi Qi[8], Qian Wang[5], Avinash Kori[6], Olivier Gevaert[7], Yunlong Zhang[8], Dinggang Shen[9,10], Mahendra Khened[6], Xinghao Ding[8], Ganapathy Krishnamurthi[6], Jayashree Kalpathy-Cramer[11‡], James Davis[12‡], Tianhao Zhao[12‡], Rajarsi Gupta[1,12‡], Joel Saltz[1‡] and Keyvan Farahani[13‡]*

[1] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, United States, [2] Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA, United States, [3] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [4] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [5] Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, [6] Department of Engineering Design, Indian Institute of Technology Madras, Chennai, India, [7] Department of Medicine and Biomedical Data Science, Stanford University, Stanford, CA, United States, [8] School of Informatics, Xiamen University, Xiamen, China, [9] Department of Radiology and BRIC, The University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [10] Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea, [11] Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States, [12] Department of Pathology, Stony Brook University, Stony Brook, NY, United States, [13] Cancer Imaging Program, National Cancer Institute, National Institutes of Health, Bethesda, MD, United States

Biomedical imaging Is an important source of information in cancer research. Characterizations of cancer morphology at onset, progression, and in response to treatment provide complementary information to that gleaned from genomics and clinical data. Accurate extraction and classification of both visual and latent image features Is an increasingly complex challenge due to the increased complexity and resolution of biomedical image data. In this paper, we present four deep learning-based image analysis methods from the Computational Precision Medicine (CPM) satellite event of the 21st International Medical Image Computing and Computer Assisted Intervention (MICCAI 2018) conference. One method Is a segmentation method designed to segment nuclei in whole slide tissue images (WSIs) of adult diffuse glioma cases. It achieved a Dice similarity coefficient of 0.868 with the CPM challenge datasets. Three methods are classification methods developed to categorize adult diffuse glioma cases into oligodendroglioma and astrocytoma classes using radiographic and histologic image data. These methods achieved accuracy values of 0.75, 0.80, and 0.90, measured as the ratio of the number of correct classifications to the number of total cases, with the challenge datasets. The evaluations of the four methods indicate that (1) carefully constructed deep learning algorithms are able to produce high accuracy in the analysis of biomedical image data and (2) the combination of radiographic with histologic image information improves classification performance.

Keywords: digital pathology, radiology, segmentation, classification, image analysis, deep learning

OPEN ACCESS

Edited by:
John Ashburner,
University College London,
United Kingdom

Reviewed by:
Houman Sotoudeh,
University of Alabama at Birmingham,
United States
Javad Noorbakhsh,
Broad Institute, United States

***Correspondence:**
Tahsin Kurc
tahsin.kurc@stonybrook.edu

† These authors have contributed equally to this work and share first authorship

‡ People involved in the organization of the challenge

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *29 August 2019*
Accepted: *10 January 2020*
Published: *21 February 2020*

Citation:
Kurc T, Bakas S, Ren X, Bagari A, Momeni A, Huang Y, Zhang L, Kumar A, Thibault M, Qi Q, Wang Q, Kori A, Gevaert O, Zhang Y, Shen D, Khened M, Ding X, Krishnamurthi G, Kalpathy-Cramer J, Davis J, Zhao T, Gupta R, Saltz J and Farahani K (2020) Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches. Front. Neurosci. 14:27. doi: 10.3389/fnins.2020.00027

INTRODUCTION

Cancer is a major life-threatening health problem around the world. More than 1.7 million new cancer cases and over 600,000 cancer deaths are estimated in 2019 in the United States alone (Siegel et al., 2019). Brain cancer is one of the deadliest cancer types with low survival rates among both women and men (Siegel et al., 2016; Yuan et al., 2016). Cancer research relies on accurate and reproducible disease characterizations in order to better understand what triggers cancer and how cancer progresses so that more effective means of evaluating cancer interventions can be developed. This requires assembling observational and experimental data at multiple biological scales and fusing information from multiple data modalities.

Biomedical imaging is one of the crucial data modalities in cancer research. Features gleaned from high-resolution, detailed images play a key role in the development of correlative and predictive representations of cancer morphology. Combined with clinical and genomics data, image features can result in more effective data-driven research and healthcare delivery for cancer patients. Biomedical imaging, hence, has evolved into an indispensable tool for researchers and clinicians to extract, analyze, and interpret the complex landscape of diagnostic and prognostic information and to assess treatment strategies. Radiology and the rapidly growing field of Radiomics provide a means of quantitative study of cancer properties at the macroscopic scale. Radiomics deals with the extraction, analysis, and interpretation of large sets of visual and sub-visual image features for organ-level quantification and classification of tumors (Lambin et al., 2012; Gillies, 2013; Aerts et al., 2014; Parmar et al., 2015; Gillies et al., 2016; Zwanenburg et al., 2016). The histopathologic examination of tissue, on the other hand, reveals the effects of cancer onset and progression at the sub-cellular level (Gurcan et al., 2009; Foran et al., 2011; Kong et al., 2011; Kothari et al., 2013; Griffin and Treanor, 2017; Yonekura et al., 2018). Histopathology has been used as a primary source of information for cancer diagnosis and prognosis. Diagnosis and grading of brain tumors, for example, is traditionally done by a neuropathologist examining stained tissue sections fixed on glass slides under a light microscope. Radiology is a more prevalent imaging modality in research and clinical settings. Advancements in digital microscopes made it possible to capture high-resolution images of whole slide tissue specimens and tissue microarrays, enabling increased use of virtual slides in histopathologic analysis.

In this paper, we present the application of state-of-the-art image analysis methods for segmentation and classification tasks for radiographic and histologic image data. We describe a collection of four deep learning-based methods: one method for the segmentation of nuclei and three methods for the classification of brain tumor cases. These methods are from the challenge teams who achieved the top scores at the Computational Precision Medicine (CPM) satellite event of the 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI2018) and agreed to contribute to this summary manuscript. The CPM event was organized by a subset of the co-authors on this paper as a cluster of image analysis challenges. It is one of the series of challenges organized since 2014 to provide a platform for biomedical imaging research teams to evaluate state-of-the-art algorithms in a controlled environment.

The 2018 CPM event targeted brain diffuse glioma and consisted of two sub-challenges. The *first sub-challenge* was designed to evaluate the performance of algorithms for the detection and segmentation of nuclear material in tissue images. We describe a nucleus segmentation method from this sub-challenge. The method employs an adaptation of the Mask-RCNN algorithm to solve the problem of cell segmentation in hematoxylin and eosin (H&E) stained tissue microscopy images. The authors of this method developed pre- and post-processing steps to further improve the performance of the algorithm. The method achieved a Dice similarity coefficient score of 0.868 when evaluated against a set of manually segmented tissue images. The *second sub-challenge* asked participants to classify lower grade glioma (LGG) cases into oligodendroglioma and astrocytoma subtypes using both radiology and histopathology images. We present three classification methods from this sub-challenge. One of the methods refines lower confidence predictions from a radiology image model by combining predictions from a tissue image model. The second method implements two distinct classification models for radiographic and histologic images and combines them through a dropout-enabled ensemble learning. The third method uses multiple deep learning models: one model for classifying tissue images and two models for segmenting and classifying radiology images. A weighted average operation is then applied to the classification results from tissue and radiology images to assign a class label to each case. The methods achieved accuracy values of 0.90, 0.80, and 0.75, respectively—accuracy was measured as the number of correctly classified cases divided by the total number of cases.

In addition to presenting these algorithms, we intend to make the datasets used in the MICCAI CPM 2018 challenge publicly available to provide a valuable resource for development and refinement of future segmentation and classification algorithms.

MATERIALS AND METHODS

In this section, we first present a brief overview of existing work on biomedical image analysis (section "Related Work"). We describe the CPM challenge and datasets in Section "Datasets and Performance Evaluation." We present the nucleus segmentation method in Section "Instance Segmentation of Nuclei in Brain Tissue Images" and the three classification methods in Section "Methods for Classification of Brain Cancer Cases."

Related Work

Computer-aided analysis and interpretation of image data is crucial to maximizing benefits from biomedical imaging. Common image analysis operations include segmentation of regions and objects (e.g., nodules and cells) and classification of image regions and images into categories. Image features

and quantitative measures obtained from segmentation and classification can be used in downstream analyses that integrate information from clinical and molecular data and develop predictive and correlative models. Studies have shown the value of image analysis and image features in research, and an increasing number of research projects have developed image analysis methods to efficiently, accurately, and reliably convert raw image data into rich information and new knowledge (Gurcan et al., 2009; Foran et al., 2011; Kong et al., 2011; Kothari et al., 2012, 2013; Lambin et al., 2012; Gillies, 2013; Cheng et al., 2016; Coroller et al., 2016; Gao et al., 2016; Ishikawa et al., 2016; Madabhushi and Lee, 2016; Manivannan et al., 2016; Xing and Yang, 2016; Al-Milaji et al., 2017; Bakas et al., 2017c; Lehrer et al., 2017; Chang et al., 2018a, 2019; Fabelo et al., 2018; Hu et al., 2018; Khosravi et al., 2018; Lee et al., 2018; Mobadersany et al., 2018; Peikari et al., 2018; Saltz et al., 2018; Yonekura et al., 2018; Zhou et al., 2018). Recent work on biomedical image analysis focused on the development and application of machine learning methods, in particular, deep learning models.

The work done by Qian et al. detected and differentiated GBM from solitary brain metastases (van Griethuysen et al., 2017) using a support vector machine (SVM) model. The analysis algorithm computes a variety of radiomic features, using the PyRadiomics package (van Griethuysen et al., 2016; Lu et al., 2019), from contrast-enhanced Radiology image datasets. The experiments show that a combination of the least absolute shrinkage and selection operator (LASSO) and SVM achieves the best prognostic prediction performance and the highest stability. Lu et al. (Krizhevsky et al., 2012) proposed and evaluated an approach, which uses the AlexNet deep learning network (Abrol et al., 2018) as a feature extractor and applies transfer learning to train a model for brain disease detection in magnetic resonance imaging (MRI) data. The last three layers of AlexNet are replaced by a fully connected layer, a softmax layer, and a classification layer to implement the feature extractor function. Chang et al. (2018a) proposed and implemented a CNN model to predict isocitrate dehydrogenase (IDH) mutations in glioma patients using preoperative MRI data. Their experimental evaluation shows that incorporating the age at which a patient was diagnosed with cancer improves algorithm accuracy to 89%. Abrol et al. (Binder et al., 2018) applied feature selection and SVM-based classification methods on MRI data obtained from a group of GBM patients. Their experimental results show that three-dimensional radiomic features computed from radiology images could be used to differentiate pseudo-progression from true cancer progression in GBM patients. Binder et al. (Shukla et al., 2017) identified radiographic signatures of extracellular domain missense mutants (i.e., A289V) of the epidermal growth factor receptor (EGFR) suggestive of an invasive and proliferative phenotype, and associated with shorter patient survival. Their approach leverages the integrated analysis of advanced multi-parametric MRI (Bakas et al., 2016) and biophysical tumor growth modeling (Akbari et al., 2018). Their findings were corroborated by experiments *in vitro* and *in vivo* in animal models, contributing to the discovery of a potential molecular target and presenting an opportunity for potential therapeutic development (Shukla et al., 2017). Another study (Bakas et al., 2017a) found an imaging signature in radiology images of the most prevalent mutation of EGFR, namely, EGFRvIII, revealing a complex yet distinct macroscopic GBM radiographic phenotype. This signature showed a classification accuracy of ∼90% for determining EGFRvIII GBM tumors. The study used an SVM model for multivariate integrative analysis of multiple image features to identify the signature. The features include the tumor's spatial distribution pattern leveraging a biophysical growth model (Akbari et al., 2018) and a distinct within-patient self-normalized heterogeneity index (Wang et al., 2019).

Mobadersany et al. (2018) examined the application of deep learning techniques to predict outcomes in LGG and glioblastoma multiforme (GBM) patients. Their approach combines tissue image analysis results with genomics data to achieve high accuracy. The deep learning network consists of convolutional layers, which are trained to predict image patterns associated with survival. This network is connected to fully connected layers that transform the image features for survival analysis. Survival data are modeled via a Cox proportional hazard layer. Wang et al. (Qian et al., 2019) implemented an analysis pipeline to classify glioma cases into grades II, III, and IV gliomas using whole slide tissue images (WSIs) from H&E and Ki-67 stained tissue samples. The pipeline consists of multiple steps, including region-of-interest (ROI) identification, image feature extraction, feature selection, automated grading of slides, and interpretation of the grading results. Multiple image features, such as the shapes and sizes of nuclei and image intensity distribution, are computed and pruned using a random forest method. The grading step employs machine learning models with automatic tuning of model parameters for the best classification performance. Saltz et al. (2018) employed a deep learning workflow to create maps of tumor-infiltrating lymphocytes (TILs) in more than 5,000 WSIs from 13 different cancer types in The Cancer Genome Atlas (TCGA) repository. The image analysis approach partitions each WSI into small (50 μm by 50 μm) patches and classifies each patch as either TIL-positive or TIL-negative. The workflow implements an iterative learning phase in which predictions by the deep learning models are reviewed and corrected by pathologists, to refine and improve classification accuracy. The analysis method also uses a convolutional neural network (CNN) to identify and segment regions of necrosis in order to reduce false positives.

Nucleus segmentation is one of the core analysis tasks in histopathology imaging projects which study tissue morphology (Gurcan et al., 2009; Madabhushi and Lee, 2016; Xing and Yang, 2016). The nucleus segmentation task is challenging because of the relatively large variation in the intensity of captured signal and the ambiguity of boundary information when separating neighboring nuclei. Several projects proposed machine learning algorithms that use engineered image features and algorithms that perform statistical analyses of intensity and texture properties to detect and delineate nucleus boundaries (Kong et al., 2011; Gao et al., 2016; Peikari and Martel, 2016; Peikari et al., 2018). In recent years, there has been a significant shift toward the application of deep learning techniques. Yang et al. (2018) proposed a method that uses a U-Net model

to segment lesions in cervical cancer cases. The segmentation results are fed into a cascade network, which integrates the foreground and the edges of the segmented nuclei to generate instance segmentations. Wollmann et al. (2019) developed a hyperparameter optimization method that searches for the best parameters of a nucleus segmentation pipeline to improve segmentation accuracy. The authors evaluated their technique with two analysis pipelines, a clustering-based pipeline and a deep learning pipeline, using prostate cancer tissue images. Their results show that the deep learning pipeline performs better than the clustering-based pipeline. Alom et al. (2018) proposed a residual recurrent CNN built on the U-Net architecture (Ronneberger et al., 2015). While this type of network has been used for segmentation of macro-level objects such as retinal blood vessels and the lungs, the authors adopted it to segmentation of the nuclei. Xie and Li (2018) implemented a neural network method that learns object-level and pixel-level information in tissue image patches. The goal is to have the analysis pipeline carry out nucleus detection and nucleus segmentation simultaneously. Hou et al. (2019b) proposed a sparse convolutional autoencoder for the detection of nuclei and feature extraction in WSIs. The approach integrates nucleus detection and feature learning in a single network. The network encodes the nuclei into sparse feature maps, which represent the nuclei's locations and appearances and can be fine tuned for end-to-end supervised learning.

Radiology and pathology capture morphologic data at different biological scales. The non-invasive and non-ionizing property of MRI made it quite popular for oncology imaging studies such as brain tumors (Bakas et al., 2016). On the other hand, the *de facto* standard for tumor assessment and grading is whole slide tissue biopsy examined under a microscope. Combined use of image modalities from both domains can lead to improvements in image-based analyses. Lundstrom et al. (2017) argue for a tighter collaboration between radiology, pathology, and genomics teams toward enhanced integrated diagnosis of disease. The authors point to the increasing use of digital slide technologies in pathology as well as to the fact that computational approaches for radiology and pathology imaging modalities are not fundamentally different. They note that combining complementary views of the disease from multiple scales can maximize the benefits of biomedical imaging. Madabhushi and Lee (2016) note that researchers are increasingly looking at opportunities for combining radiomic data with features extracted from high-resolution pathology image for better predictive capabilities in disease prognosis. On the methodology and software front, Arnold et al. (2016) developed a web-based platform that integrates radiology and pathology data for cancer diagnosis. Saltz et al. (2017) devised methods and tools for combined computation, management, and exploration of image features from radiology and pathology image datasets. Kelahan et al. (2017) implemented a dashboard for radiologists to view pathology reports to aid with diagnosis and image-guided decision making. McGarry et al. (2018) proposed a method for combining multi-parametric MRI data with digital pathology slides to train predictive models for prostate cancer localization.

Despite a growing body of research and development on methods and tools, computerized image analysis continues to be a challenging task. Both image resolutions and data complexity continue to increase, requiring the enhancement of existing methods and the development of new techniques. For example, contemporary digital microscopy scanners are capable of imaging whole slide tissue specimens at very high resolutions (e.g., over 80,000 × 80,000 pixels). These images may contain millions of cells and nuclei, and multiple types of regions (e.g., tumor, stromal, and normal tissues). There can be significant morphological heterogeneity within a specimen, as well as across specimens in both radiographic and histologic imaging, requiring novel methods that can handle heterogeneity and increasing the density of morphologic information.

Datasets and Performance Evaluation

The approaches, which will be described in Sections "Instance Segmentation of Nuclei in Brain Tissue Images" and "Methods for Classification of Brain Cancer Cases," were experimentally evaluated with radiographic and histologic image datasets from the MICCAI 2018 CPM challenge event. Here we provide a brief description of the challenge datasets and the methods for scoring algorithm performance. The datasets for the 2018 CPM challenge were obtained from TCGA[1] (Tomczak et al., 2015) and The Cancer Imaging Archive (TCIA[2]) (Clark et al., 2013; Prior et al., 2013) repositories, and the images had been scanned at the highest resolution. Images from these sources are publicly available and have been used in many publications (e.g., Aerts et al., 2014; Yu et al., 2016; Bakas et al., 2017c; Mobadersany et al., 2018; Saltz et al., 2018; Agarwal et al., 2019).

Datasets for Segmentation of Nuclei in Pathology Images

A WSI may contain hundreds of thousands of nuclei; some images with large tissue coverage will have more than one million nuclei. Manually segmenting all nuclei in the entire WSIs would be infeasible. Thus, we extracted image tiles from WSIs and used the tiles in the training and test datasets in order to reduce the cost of generating high-quality ground truth data as well as the computational requirements of the training and test steps of analysis algorithms. The image tiles were selected by a pathologist and extracted from a set of GBM and LGG WSIs at the highest resolution. The training and test datasets consisted of 15 and 18 image tiles, respectively. The sizes of the tiles ranged from 459 × 392 pixels to 1032 × 808 pixels in the training set and from 378 × 322 pixels to 500 × 500 pixels in the test set. The nuclei in each image tile were segmented by two students. The segmentation results were reviewed, refined, and consolidated by the pathologist to generate the final set of segmentation data. This process generated 2905 and 2235 nuclei in the training and test sets, respectively.

In the challenge event, the performance of a segmentation algorithm was measured as the average of the standard Dice similarity coefficient and a modified version of the Dice metric.

[1] https://portal.gdc.cancer.gov

[2] https://www.cancerimagingarchive.net

The standard Dice score (Dice, 1945) measures the overlap between two sets of segmentation results without taking into account the individual nuclei. That is, it computes the amount overlap between the ground truth mask and the mask generated by the segmentation algorithm without considering splitting and merging of the nuclei by the algorithm. The modified Dice metric aims to incorporate split and merge errors into the score. We refer the reader to an earlier publication (Vu et al., 2019) for a more detailed description of the modified Dice metric.

Datasets for Combined Radiology and Pathology Classification

The datasets were matched MRI and digital pathology images obtained from the same patients and the same time point. Each case corresponded to a single patient. There was one set of MRI data (T1, T1C, FLAIR, and T2 images) and one corresponding WSI for each case. The training set contained a total of 32 cases: 16 cases that were classified as oligodendroglioma and 16 cases classified as astrocytoma. The test dataset consisted of 20 cases with 10 cases of oligodendroglioma and 10 cases of astrocytoma. We retrieved the WSI and MRI images from the TCGA and TCIA archives, respectively. These images had been obtained and classified following the protocols implemented in the TCGA project[3]. We obtained the ground truth classification labels of the cases from the associated clinical and metadata in the TCGA repository. These classifications were further reviewed by a pathologist and a radiologist. In the challenge event, we used the accuracy of a classification method to score its performance and rank it. We counted the number of correctly classified cases and divided that number by the total number of cases to compute the accuracy score.

In the following sections, we will present a nucleus segmentation algorithm (section "Instance Segmentation of Nuclei in Brain Tissue Images"), which achieved the second highest score in the segmentation challenge, and three classification algorithms (section "Methods for Classification of Brain Cancer Cases"), which achieved the top three scores in the classification challenge.

Instance Segmentation of Nuclei in Brain Tissue Images

In this section, we present the nucleus segmentation algorithm developed by XR, QW, LZ, and DS. This method achieved the second highest score in the CPM challenge and its developers agreed to contribute to this manuscript.

The method implements an application of the Mask-RCNN network (He et al., 2017) with a novel MASK non-maximum suppression (MASK-NMS) module, which can increase the robustness of the model. Mask-RCNN is a deep learning network extended from the Faster-RCNN model (Ren et al., 2015) and is used to carry out semantic and object instance segmentation (see **Figure 1**). In our implementation, we used ResNet-101 to build a Mask-RCNN pyramid network backbone for the segmentation of nuclei in WSIs. This adaptation is based on an existing implementation by Matterport[4]. We have extended this implementation in several ways to improve segmentation performance. First, we have reduced the region proposal network (RPN) anchor sizes and increased the number of anchors to be used because the nuclei are small objects and can be found anywhere in a tissue image. Second, we have increased the maximum number of predicted objects, since even a small image tile from a tissue slide can contain 1000 or more nuclei. Moreover, rather than training the network end-to-end from the start, we initialized the model using weights from the pre-training on the MSCOCO dataset (Lin et al., 2014). We train the layers in multiple stages. We first train the network heads after they are randomly initialized. We later train the upper layers of the network. After this, we reduce the learning rate by a factor of 10 and train the entire network end to end. In our experiments, the training took 300 epochs using stochastic gradient descent with momentum set to 0.9. During training and testing, input tissue images were cropped to 600×600.

In addition to the above extensions, we implemented a set of pre-processing steps to further improve the algorithm performance. Holes in the masks are filled by an image morphology operation. Fused nucleic masks are split by applying morphological erosion and dilation. To help avoid overfitting, data augmentation, which could increase the amount of training data, is applied in the form of random crops, random rotations, Gaussian blurring, and random horizontal and vertical flips.

Our implementation combines predictions from fivefold cross training models in a post-processing step (see **Figure 1**). We have implemented this step in a novel module called MASK-NMS, which is one of our contributions in the segmentation method. MASK-NMS takes unions of masks with maximum overlap and removes false-positive masks with a small overlap. It starts with a set of segmentation results. This set is called I. Each result in set I is assigned a score S, which is the value of the classification probability from the Mask-RCNN module and corresponds to the confidence level of the segmentation result. After selecting the segmentation with the maximum score M (the maximum score among scores S), MASK-NMS removes it from the set I and appends it to the final segmentation set D. D is initialized to an empty set. It also removes any segmentations with an overlap greater than a threshold N in the set I, where the intersection over union (IOU) is used as the overlap metric. IOU is also known as the Jaccard similarity index (Jaccard, 1901), which measures the similarity between finite sample sets. It is defined as the size of the intersection between two sets divided by the size of the union of the sets. The selection process repeats until set I becomes empty. Finally, we obtain the segmentation results in set D. The MASK-NMS module assembles multiple results together and reduces false positives and false negatives.

Methods for Classification of Brain Cancer Cases

In this section, we present three classification algorithms, which achieved the top three scores in the classification challenge and the developers of which agreed to contribute to this manuscript.

[3]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/types

[4]https://github.com/matterport/Mask_RCNN

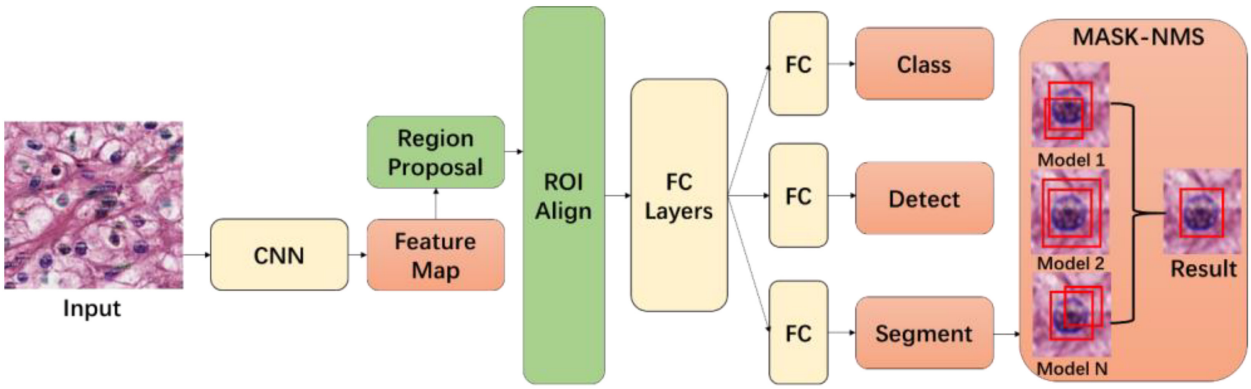


FIGURE 1 | Tissue image segmentation model. The first part of the model consists of the Mask-RCNN module. Output from this module is input to the MASK-NMS module for final segmentation prediction output.

An Approach for Classification of Low-Grade Gliomas Using Combined Radiology and Pathology Image Data

The top-performing method (developed by AB, AsK, AvK, MK, and GK) (Bagari et al., 2018) in the classification challenge uses an MRI classification model and a WSI classification model and combines the predictions from the two models to assign a class to a given case. The overall analysis pipeline is depicted in **Figures 2–4** and described below.

Radiology image analysis pipeline

Different pulse sequences in MRI, including native T1- and T2-weighted, T2-Flair, and T1-weighted post-contrast imaging, can be used to enhance different parts of a tumor. In this part of the pipeline (**Figure 2**), we execute a segmentation pipeline consisting of the following steps on these images before features are computed from the images and used in the classification model: (1) Skull stripping: It is necessary to remove the skull from MRI as its presence can be wrongly interpreted as a tumor, and most segmentation networks are trained using skull-stripped images. (2) Co-registration and re-sampling to isotropic voxel spacing: Following skull stripping is the step of co-registering the MRI sequences to a reference sequence. Generally, there can be movement between scans if the patient does not remain still or if the scan is acquired on a different day or using a different scanner. Registered images are spatially correlated across channels and can be used for tumor segmentation. We register sequences T1, FLAIR, and T2 with respect to T1c scan. The MRI volumes are re-sampled to an isotropic voxel resolution of 1 mm^3 after the co-registration step. (3) Segmentation of tumor regions using a CNN: Tumor regions are segmented by a fully CNN trained on the BraTS-2018 dataset (Menze et al., 2014; Bakas et al., 2017b,c, 2018; Crimi et al., 2018). After the segmentation step, a set of 105 radiomic features are computed on segmented regions using the pyradiomic library (Lu et al., 2019). These features include shape features, first-order statistics, features from gray level co-occurrence matrix, features from gray level run length matrix and gray level size zero matrix, and neighboring gray tone difference matrix. The 105-dimensional radiomic feature vectors are reduced to a 16-dimensional feature vector using the principal component analysis. A classification model is trained with 16-dimensional feature vectors as input. If the training dataset has N cases, the model is trained with an (*N*,16) input using logistic regression with the liblinear optimization algorithm (Fan et al., 2008) and a fivefold cross-validation process. This process fits a logistic regression model on the entire training data. Classification predictions from the MRI data are obtained using this model.

Analysis pipeline for whole slide tissue images

Tissue slides may contain large areas of glass background that are irrelevant to image analysis and should be removed. In this part of the pipeline (**Figure 3**), in order to detect and segment tissue regions and remove regions corresponding to glass background, a tissue image is first converted from the RGB color space to the HSV color space. Then, lower and upper thresholds are applied on color intensities to get a binary mask. The binary mask is processed to fill in small holes and remove clustered clumps from foreground pixels. After this step, bounding boxes around all the discrete contours are obtained. The bounding boxes serve as blueprints for the patch extraction process. The patch extraction process partitions the segmented tissue region into 224 × 224-pixel patches. The 224 × 224-pixel patches are color-normalized (Reinhard et al., 2001) and assigned the same label as the label of the WSI. A subset of distinct patches is filtered out using an outlier detection technique called the Isolation Forest (Liu et al., 2008). The filtering step is executed as follows. We train an autoencoder with a pixel-wise reconstruction loss to generate feature vector representations of patches from the input image. The isolation forest method is then executed with these feature vectors to find outlier patches. The remaining patches after the outlier detection step are used to refine a DenseNet-161 network, which has been pre-trained on ImageNet. Binary cross entropy is used as the loss function. During the prediction phase, test patches extracted from a WSI are classified using the trained model, and a probability score is assigned to the image based on a voting of classes predicted for individual patches.

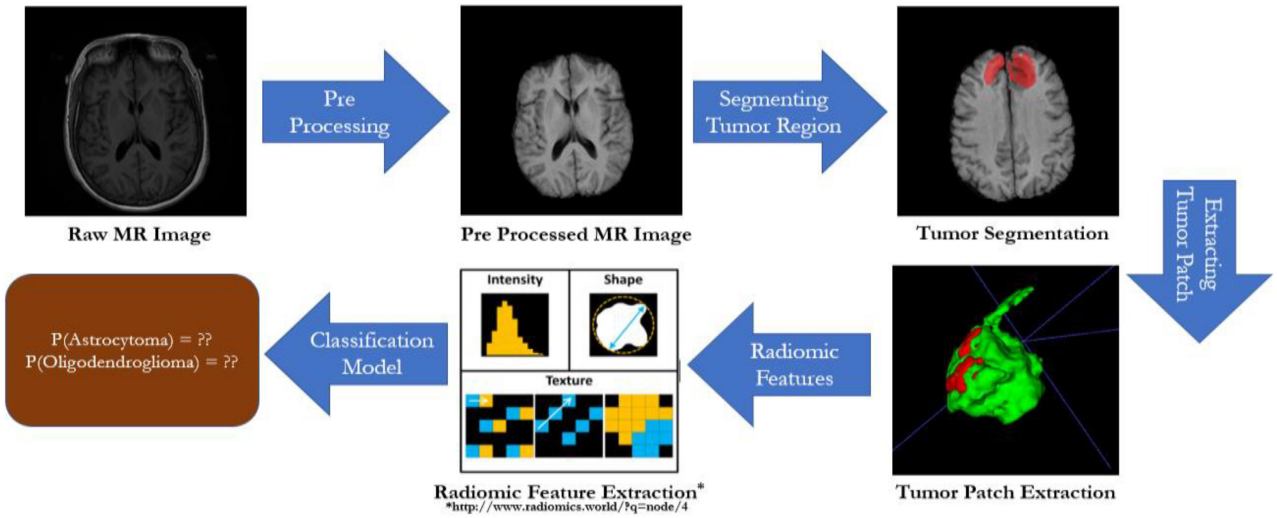


FIGURE 2 | Radiology image analysis. Images are pre-processed (i.e., skull stripping and co-registration) before they are analyzed through the remaining steps of the analysis pipeline. After the pre-processing step, tumor regions in the images are segmented via a CNN model. This step is followed by computation of a set of 105 radiomic features in segmented regions. The high-dimensional feature vector is reduced to a 16-dimensional feature vector using the principle component analysis method. A classification network is trained with these feature vectors.

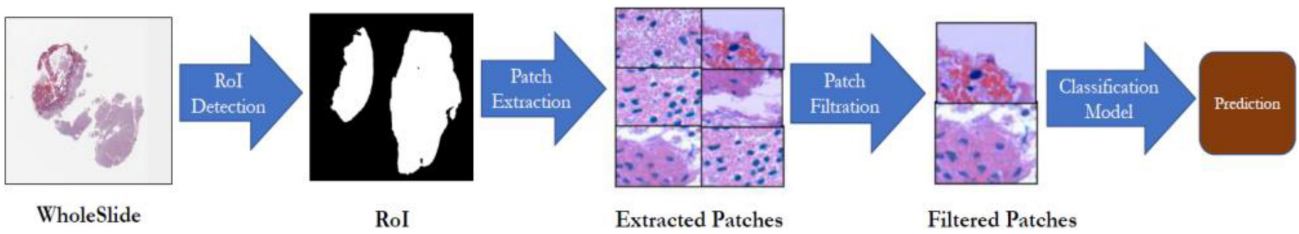


FIGURE 3 | Pathology image analysis. A region-of-interest (ROI) step detects and segments tissue regions. The tissue regions are partitioned into patches. Distinct patches are filtered using the isolation forest technique. The prediction represents the probability values of the case being astrocytoma or oligodendroglioma.

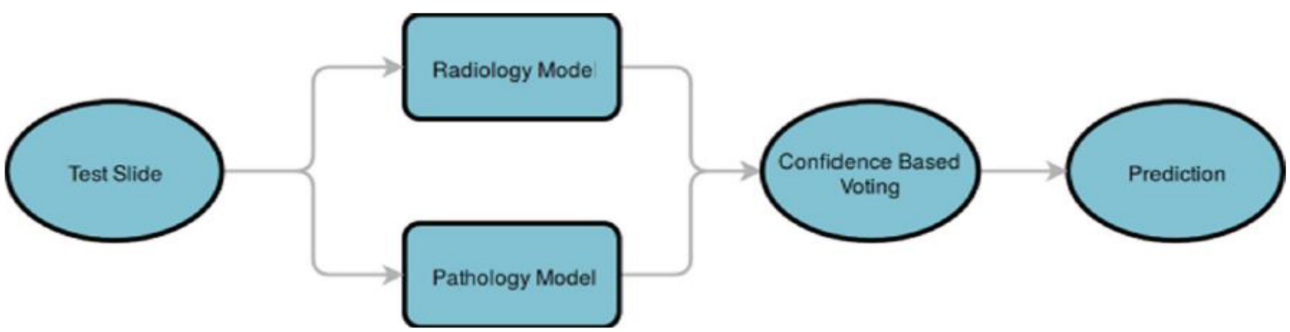


FIGURE 4 | Combining predictions from the pathology and radiology models. A test case is analyzed by the radiology classification model and the pathology classification model. The results from the two models are processed in a confidence-based voting step, which chooses the class with the highest prediction probability value.

Combining predictions

As is shown in **Figure 4**, finally, predictions from both the radiology and pathology models are compared, and the class label of a case is determined based on the model, which gives a prediction with a higher probability score.

Dropout-Enabled Ensemble Learning for Multi-Scale Biomedical Image Classification

This method is the second best performing (developed by AM, MT, and OG) (Momeni et al., 2018) and proposes two distinct classification models for radiographic and histopathologic

images and their integration through dropout-enabled ensemble learning.

Radiology classification model

As is shown in **Figure 5**, radiology images are pre-processed through a pipeline of bias field correction, skull-stripping, and co-registration steps before they are input to a 3D CNN network. The 3D CNN consists of eight layers to extract deep features from MRI and three 3D max pooling to reduce the sample size. The input of the 3D CNN is a 3D voxel image in the form of three spatial dimensions and two modalities per voxel. In this work, the 3D CNN is trained with the T1c and T2-FLAIR modalities only because these are the most informative for LGG segmentation. After the last convolutional layer, the extracted features are averaged over all of the 3D space to yield a unique 100-dimensional feature vector per case. This vector is connected to a 1D output for classification with cross-entropy loss. The whole network is then trained. To avoid overfitting, we use classical data augmentation techniques (rotation, cropping, etc.) as well as dropout. Eight dropout layers are placed throughout the network to avoid overfitting and for the ensemble learning step.

Histopathology classification model

A multiple instance learning approach, as shown in **Figure 6**, is implemented for the histopathology images. The learning step is carried out after a pre-processing phase. The pre-processing steps here consist of tissue detection, color normalization, and tiling. Tissue detection is done with Otsu thresholding to detect and segment tissue regions only, eliminating regions that are glass background. A simple histogram equalization algorithm is used for color normalization prior to tiling. The tiling step extracts 20 448 × 448-pixel patches from a WSI by uniform random sampling. Once the image patches have been extracted, a DenseNet network pretrained on ImageNet is fine tuned, after removing its last fully connected layer. The remainder of the network is used as a fixed feature extractor for tissue images, and two fully connected layers with dropout are used for classification. As with the radiology model, we used

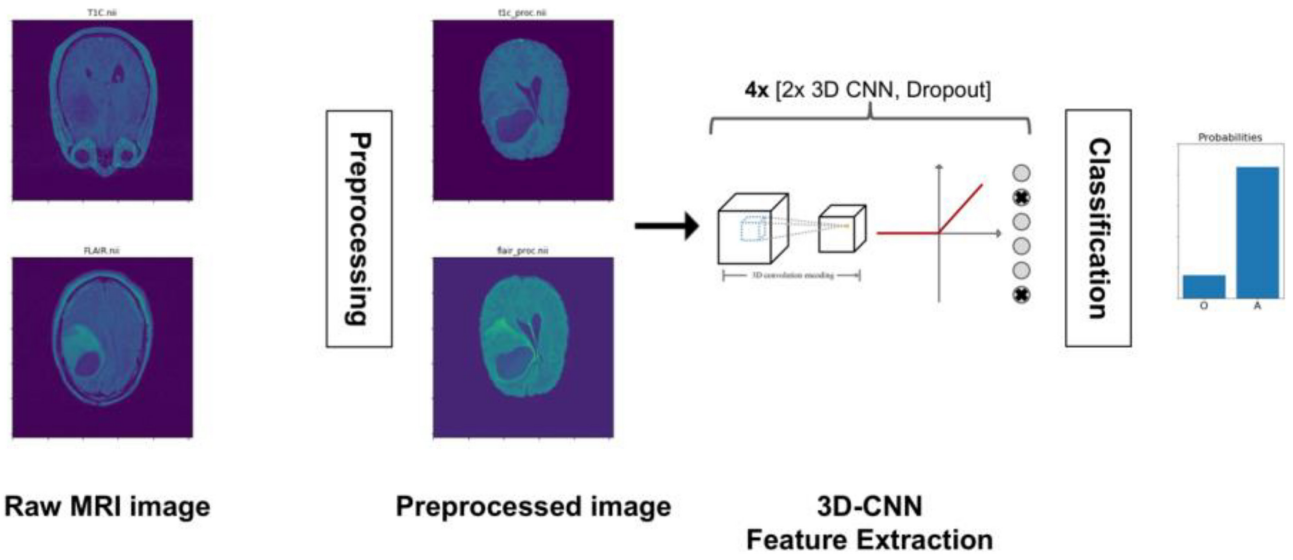


FIGURE 5 | Radiology image analysis pipeline. Radiology images are pre-processed for bias field correction, skull stripping, and co-registration before they are input to a 3D CNN. The 3D CNN is trained to output a prediction (probability) value for each case as to whether the case is oligodendroglioma (O) or astrocytoma (A).

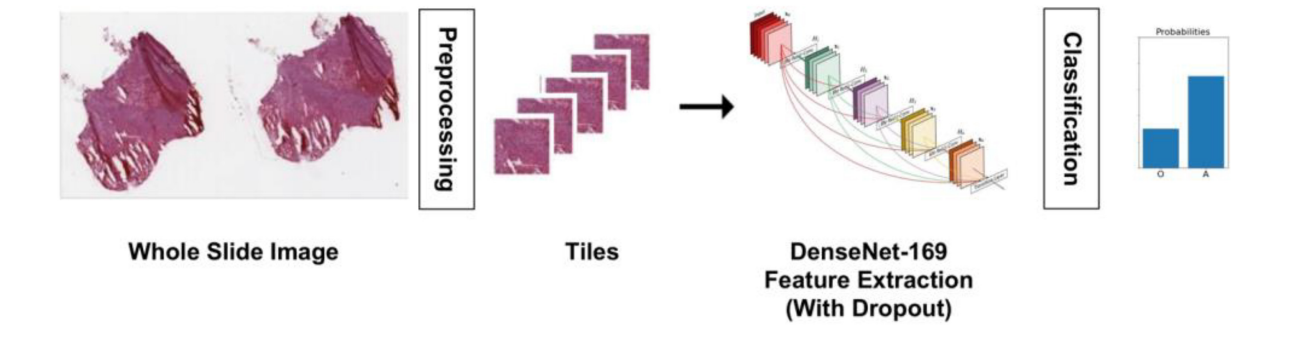


FIGURE 6 | Histopathology image analysis pipeline. The whole slide tissue images are pre-processed to detect tissue, do color normalization, and extract tiles. The tiles are input to a DenseNet model for classification. The model outputs the probability of a case being oligodendroglioma or astrocytoma.

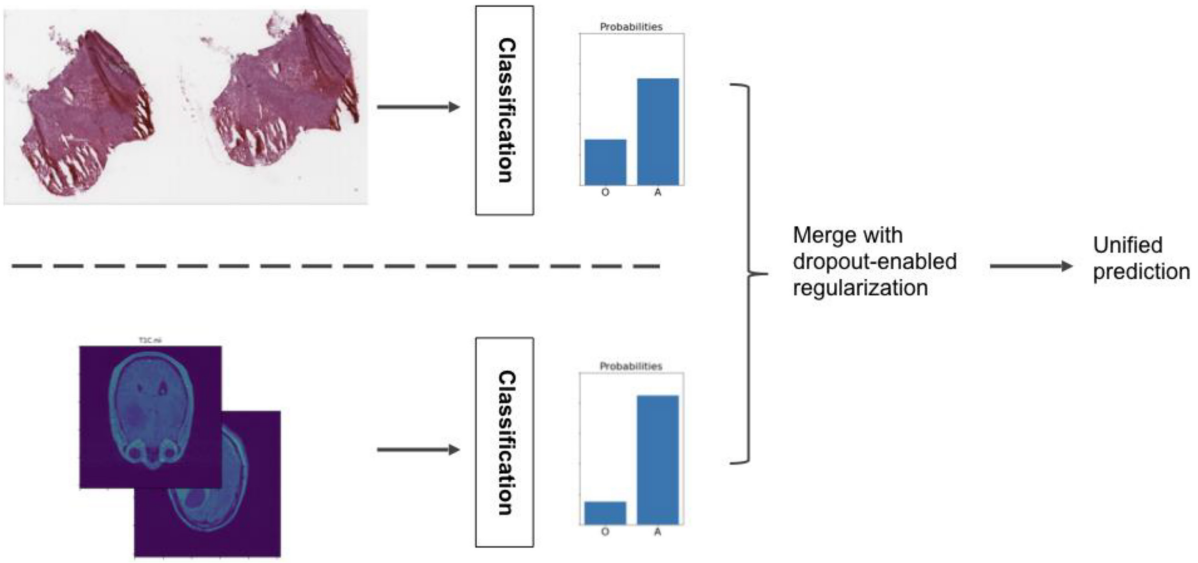


FIGURE 7 | Ensemble model that combines classifications from the radiology and histopathology image analysis pipelines.

classical data augmentation techniques along with dropout to eliminate overfitting.

Ensemble learning model

The main contribution of our approach is a meta-algorithm that combines the histopathology and radiology classification models, as is shown in **Figure 7**. In this ensemble learning methodology, each model is trained separately. Their predictions are combined into a single, more robust output. The basic idea is to extract the one-to-last feature layer from each individual classification model and form a single feature vector for each case/patient by concatenating the two feature vectors. An SVM model is then trained with the combined feature vectors to classify the cases. However, if the training dataset is small (which is the case with the CPM challenge dataset; we have 50-dimensional feature vectors from both the classification models and only 32 cases in the training dataset), the classification problem can become under-determined and result in overfitting of the models. To address this problem, we use regularization through dropout in the ensemble learning step. The idea is to enable the dropout values of the models in the test phase, so that individual models produce multiple (typically thousands) feature vectors for each subject. These many feature vectors can then be concatenated to form the combined feature vectors, creating a training dataset big enough for the SVM model (Momeni et al., 2018). Dropout at test time results in sampled feature vectors that are both distinct and informative and provides sufficient variance in the training dataset. Hence, the ensemble learning method can learn a more accurate and robust model from the newly produced dataset.

A Weighted Average-Based Classification Method

The third best performing method (developed by QQ, YZ, YH, and XD) is illustrated in **Figure 8**. It analyzes each imaging modality (radiology images and pathology images) separately and combines the prediction results via a weighted average operation. We describe the individual classification models and weighted average operation below.

Classification of pathology images is carried out by identifying tissue characteristics that differentiate oligodendroglioma from astrocytoma. Astrocytoma is noted to have more grades, as well as necrosis, increased cell density, calcification, and nuclear atypia. On the other hand, fried egg-like cells, and the tissue characteristics of chicken-cage-like blood vessels are unique to oligodendroglioma. In the proposed method, each histologic image is partitioned into 512 × 512 patches. A sample set is created to identify typical samples of both subtypes of brain diffuse gliomas to assess imbalance in the data. In order to prevent the classification error caused by data imbalance, our method expands the sample set by rotating the original image in symmetrical and asymmetrical directions. The balanced samples are then sent to a CNN classifier network, which is trained to fully recognize the tissue and cell characteristics of oligodendroglioma and astrocytoma (see **Figure 8**). The method uses the VGG16 CNN network (Simonyan and Zisserman, 2014). We use data augmentation and add dropout layers or batch normalization layers to the classification model to reduce the risk of overfitting the model.

The classification model for radiology images is shown in **Figure 9**. Radiology images are pre-processed using methods from the SPM12 software (Penny et al., 2011). The methods include *Realign, Estimate, and Re-slice* to register data of the same modality in different cases; *Co-register and Estimate and Re-slice* to register different modal data of the same case; and *Segment and ImCalc* to extract the intracranial cavity. The pre-processed images are then segmented using the U-Net (Ronneberger et al., 2015) segmentation network. Patches with tumor, which are predicted by the segmentation network, are used as training data for a 2D Densenet (Huang et al., 2017) network. We classify each

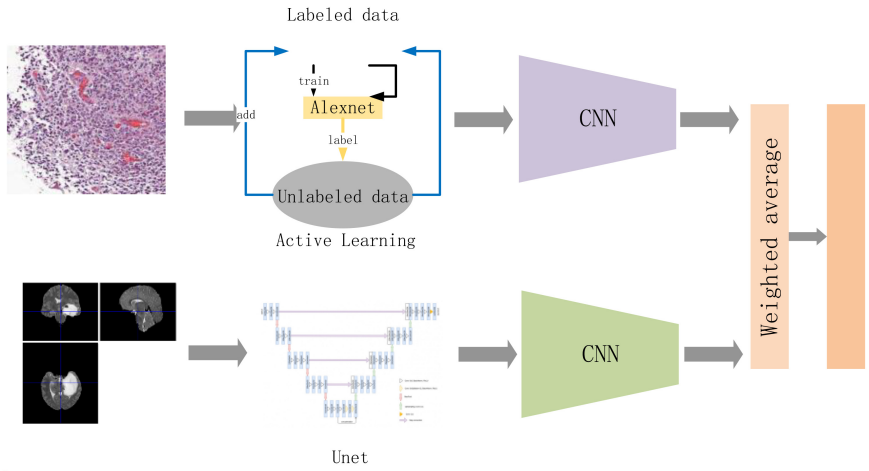


FIGURE 8 | The flow of the entire method. Among them, we slice the entire pathological data and extract the effective diseased area as much as possible. The active learning strategy follows our work in Qi et al. (2018). The goal of that work is to maximize learning accuracy from very limited labeling data. The classification model is updated iteratively with an increasing training set. The sliced pathological data are sent to a convolutional neural network to obtain the discrimination results of the pathological data. The radiological data are sent to the Unet and CNN to obtain classification results after preprocessing. Finally, the results are combined via a weighted average operation to obtain the final result.

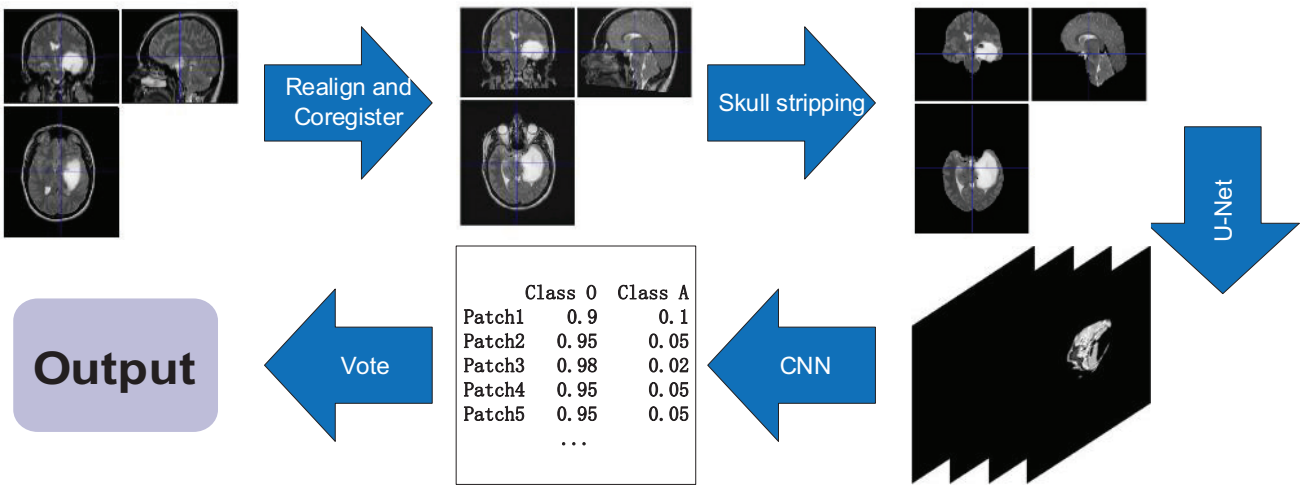


FIGURE 9 | The classification process of radiology images. The process aligns the images of different modalities through realignment and co-register, extracts brain tissue through skull stripping, extracts lesion area by the U-Net, and classifies cases by CNN.

patch, set the threshold value of 0.99, and select effective patches. The ensemble of multiple patches can effectively improve the robustness of the classifier.

Classification results from the radiology image dataset and the pathology image dataset are combined via a weighted average operation (see **Figure 8**):

$$\hat{y} = \alpha^* f(X_p) + (1 - \alpha)^* g(X_r)$$

where the classifiers for pathology data and radiology data, X_p and X_r are VGG16 and DenseNet, respectively. $f(\bullet)$ and $g(\bullet)$ represent the probabilities acquired from softmax function in X_p and X_r. The weight α is empirically estimated in predicting the final classification label $\hat{y}$.

EXPERIMENTAL RESULTS

Segmentation of Nuclei

The Mask-RCNN model with ResNet-101 backbone obtained the 45.02% mean IOU (mIOU) on fivefold validation dataset. mIOU is the average precision score for each IOU with different thresholds (from 0.05 to 0.95 in the challenge). Tissue images with nuclei detections and segmentations are illustrated in **Figure 10**. A Dice score of 0.868 was achieved with the test dataset.

Classification of Cancer Cases

On testing the algorithms on a dataset containing 20 radiology and pathology images, the three methods in Section "Methods

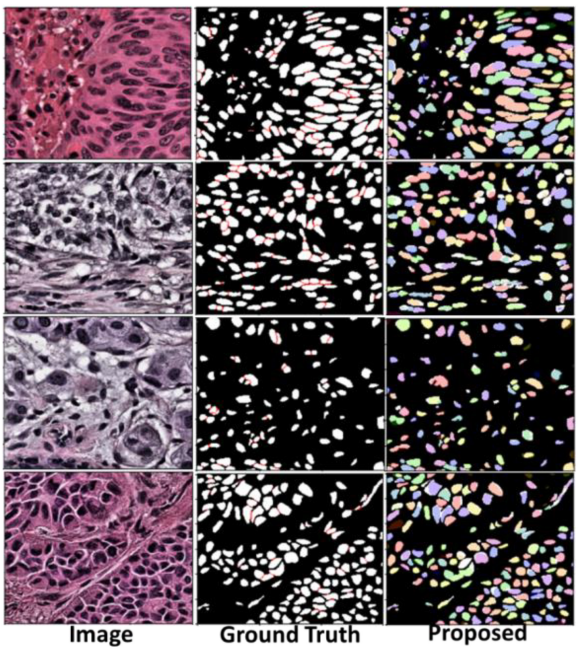


FIGURE 10 | Segmentation result in validation dataset. The **left column** shows tissue images. The **middle column** is ground truth masks. The **right column** shows results from the segmentation method.

for Classification of Brain Cancer Cases" achieved the accuracy scores (i.e., the number of correctly classified cases divided by the total number of cases) as shown in **Table 1**.

TABLE 1 | Accuracy scores of the classification methods presented in Section "Methods for Classification of Brain Cancer Cases."

Method	Score
Section "An Approach for Classification Of Low-Grade Gliomas Using Combined Radiology and Pathology Image Data"	0.90
Section "Dropout-Enabled Ensemble Learning for Multi-Scale Biomedical Image Classification"	0.80
Section "A Weighted Average-Based Classification Method"	0.75

DISCUSSION AND CONCLUSION

Biomedical imaging has made great strides in image resolution and image capture speeds over the past decade. Radiology has enjoyed a widespread adoption for many years in both research and clinical settings. New imaging technologies are now allowing researchers to capture larger volumes of more detailed radiology data. Digital microscopy scanners were emerging technologies about 20 years ago. They required constant attention to capture sharp images of tissue and took many hours to scan a tissue specimen at moderate magnification levels. Nowadays, hundreds of slide tissues can be automatically imaged in several minutes. New scanning technologies and tissue staining methods are enabling researchers to capture richer morphological information at unprecedented resolutions. We anticipate that the FDA's approval in 2017 of WSIs as a primary diagnostic tool will fuel a rapid increase in adoption of virtual slide technologies by researchers and clinicians. Combined with cheaper storage space, more powerful computing capabilities (via multi-core CPUs and accelerators such as graphics processing units), and Cloud computing infrastructures, biomedical imaging is rapidly becoming an essential tool in cancer research.

On the image analysis front, deep learning methods have seen a tremendous intake from the imaging community. These methods have demonstrated excellent results in the analysis of natural images. A rapidly growing collection of efforts are adapting these methods and extending them in innovative ways for application in biomedical image analysis. The segmentation method presented in this work shows the use of Mask-RCNN along with a non-maximum suppression (NMS) module for robust segmentation of nuclei in WSIs. The image classification methods employ a variety of deep learning methods and combine information from both radiology and pathology images to improve classification accuracy. All the methods described in this paper were evaluated with image ground truth data generated in the MICCAI CPM 2018 challenge (organized by a subset of the co-authors as denoted in the author list). The experimental results for nucleus segmentation show that high performance (i.e., high Dice scores) can be achieved by integrated use of Mask-RCNN and NMS for nucleus segmentation. The results for the classification methods show that a carefully assembled set of pipelines for each imaging modality and combination of prediction results from individual models can produce high classification accuracy.

While our work and works by other research teams have shown significant progress with more accurate, efficient, and robust image analysis algorithms, there remain challenges. One of the major challenges in machine learning analysis of biomedical imaging data is the lack of large curated and annotated training datasets, primarily because of time effort and domain expertise required for manual segmentations and classifications of tissue regions and micro-anatomic structures, such as nuclei and cells, as well as because of privacy and ownership concerns of source datasets. Some initial studies in the field of distributed learning in medicine attempted to address the data privacy and ownership challenge (Chang et al., 2018b; Sheller et al., 2018). These approaches need more investigation and adoption to facilitate collaboration across multiple medical institutions. Some projects have looked at the use of synthetic training datasets. Mahmood et al. (2018), for example, devised a method based on a conditional generative adversarial network (GAN) to improve deep learning-based segmentation of nuclei. Their method trains segmentation models using synthetic and real data. The authors employed a cycle GAN method to generate pairs of synthetic image patches and segmentation masks with varying amounts of touching and clumped nuclei. Such nuclei are difficult to segment by automated algorithms. In another work, Hou et al. (2019a) proposed a GAN architecture for the generation of synthetic tissue images and segmentation masks. The GAN architecture consists of multiple CNNs; a set of CNNs generates and refines synthetic images and masks to reference styles, and

another CNN is trained online with these images and masks to generate a segmentation model. Another GAN approach was proposed by Senaras et al. (2018b) for tumor grading. The GAN network generates synthetic image datasets with known amounts of positive and negative nuclei in immunohistochemistry-stained tissue specimens (Senaras et al., 2018b).

Another major challenge in automated biomedical image analysis is the quality assessment of input datasets and analysis results. This also is a time-consuming and labor-intensive task, as automated algorithms can process large numbers of images and generate large volumes of analysis output to be reviewed and validated, thanks to advances in computing systems. There is a need to automate the quality assessment and validation processes. Some projects are looking at this problem. A recent work by Senaras et al. (2018a) used deep learning methods to detect out-of-focus regions in WSIs so that image analysis pipelines can avoid such regions. An approach proposed by Wen et al. utilized multiple machine learning methods, namely, SVM, random forest, and CNN, to assess the quality of nuclear segmentation results. The proposed approach made use of texture and intensity features extracted from image patches in a WSI to train the quality control models (Wen et al., 2017, 2018).

As our capability to capture complex radiology and pathology image data more rapidly and at higher resolutions evolves, manual training data generation and quality evaluation will become increasingly infeasible. We expect that (semi-)automated approaches, for training data generation, for assessing the quality of data and analysis results, and for iterative refinement of deep learning models, will become important tools in a researcher's and clinician's imaging toolset. We also believe image analysis challenges, such as the MICCAI 2018 CPM challenge, are important in efforts to develop more robust methods for image analysis and method assessment and validation. One of the issues that face machine/deep learning algorithm developers is the limited amount of ground truth datasets in biomedical imaging—the small dataset size is a limitation in our work as well. Thus, in addition to providing a platform for researchers to evaluate their methods in a controlled environment, image analysis challenge events contribute to a growing set of curated datasets that are valuable resources for development and refinement of future segmentation and classification algorithms. As part of our work, we make the datasets used in this challenge available to other researchers upon request.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in The Cancer Genome Atlas (https://portal.gdc.cancer.gov) and The Cancer Imaging Archive repositories (https://www.cancerimagingarchive.net). Please contact the corresponding author for details about the datasets and how to obtain them.

AUTHOR CONTRIBUTIONS

TK, SB, JK-C, JS, and KF were the main organizers of the MICCAI 2018 CPM satellite event presented in Section "Segmentation of Nuclei." JD, TZ, RG, SB, and TK generated the datasets used in the challenge, supervised the collection of ground truth data, and developed the methods for evaluation of the challenge submissions. XR, LZ, QW, and DS proposed and implemented the nucleus segmentation algorithm presented in Section "Related Work." YH, QQ, YZ, and XD proposed and implemented the classification algorithm described in Section "Datasets for Segmentation of Nuclei in Pathology Images." AB, AsK, AvK, MK, and GK proposed and implemented the classification algorithm described in Section "Datasets for Combined Radiology and Pathology Classification." All of the authors contributed text to the paper and edited it.

FUNDING

This work was supported in part by the National Institutes of Health under award numbers NCI:U24CA180924, NCI:U24CA215109, NCI:UG3CA225021, NCI:U24CA189523, NINDS:R01NS042645, and R01LM011119 and R01LM009239 from the U.S. National Library of Medicine. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH.

REFERENCES

Abrol, S., Kotrotsou, A., Hassan, A., Elshafeey, N., Idris, T., Manohar, N., et al. (2018). Abstract 3040: radiomics discriminates pseudo-progression from true progression in glioblastoma patients: a large-scale multi-institutional study. *Cancer Res.* 78:3040.

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* 5:4006. doi: 10.1038/ncomms5006

Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., and Martí, R. (2019). Automatic mass detection in mammograms using deep convolutional neural networks. *J. Med. Imaging* 6:031409.

Akbari, H., Bakas, S., Pisapia, J. M., Nasrallah, M. P., Rozycki, M., Martinez-Lage, M., et al. (2018). In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature. *Neuro Oncol.* 20, 1068–1079. doi: 10.1093/neuonc/noy033

Al-Milaji, Z., Ersoy, I., Hafiane, A., Palaniappan, K., and Bunyak, F. (2017). Integrating segmentation with deep learning for enhanced classification of epithelial and stromal tissues in H&E images. *Pattern Recognit. Lett.* 119, 214–221. doi: 10.1016/j.patrec.2017.09.015

Alom, M. Z., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). "Nuclei Segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net)," in *Proceedings of the NAECON 2018 - IEEE National Aerospace and Electronics Conference*, Dayton, OH, 228–233.

Arnold, C. W., Wallace, W. D., Chen, S., Oh, A., Abtin, F., Genshaft, S., et al. (2016). RadPath: a web-based system for integrating and correlating radiology and pathology findings during cancer diagnosis. *Acad. Radiol.* 23, 90–100. doi: 10.1016/j.acra.2015.09.009

Bagari, A., Kumar, A., Kori, A., Khened, M., and Krishnamurthi, G. (2018). "A combined radio-histological approach for classification of low grade gliomas," in *Proceedings of the International MICCAI Brainlesion Workshop*, (Berlin: Springer), 416–427. doi: 10.1007/978-3-030-11723-8_42

Bakas, S., Akbari, H., Pisapia, J., Martinez-Lage, M., Rozycki, M., Rathore, S., et al. (2017a). In Vivo detection of EGFRvIII in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep Peritumoral infiltration: the phi-Index. *Clin. Cancer Res.* 23, 4724–4734. doi: 10.1158/1078-0432.CCR-16-1871

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. *Cancer Imaging Arch.* 2017:286. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv [preprint]*

Bakas, S., Zeng, K., Sotiras, A., Rathore, S., Akbari, H., Gaonkar, B., et al. (2016). GLISTRboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. *Brainlesion* 9556, 144–155. doi: 10.1007/978-3-319-30858-6_1

Binder, Z. A., Thorne, A. H., Bakas, S., Wileyto, E. P., Bilello, M., Akbari, H., et al. (2018). Epidermal growth factor receptor extracellular domain mutations in glioblastoma present opportunities for clinical imaging and therapeutic development. *Cancer Cell* 34:163-177.e7. doi: 10.1016/j.ccell.2018.06.006

Chang, H. Y., Jung, C. K., Woo, J. I., Lee, S., Cho, J., Kim, S. W., et al. (2019). Artificial Intelligence in Pathology. *J. Pathol. Transl. Med.* 53, 1–12.

Chang, K., Bai, H. X., Zhou, H., Su, C., Bi, W. L., Agbodza, E., et al. (2018a). Residual convolutional neural network for the determination of idh status in low- and high-grade gliomas from MR imaging. *Clin. Cancer Res.* 24, 1073–1081. doi: 10.1158/1078-0432.CCR-17-2236

Chang, K., Balachandar, N., Lam, C., Yi, D., Brown, J., Beers, A., et al. (2018b). Distributed deep learning networks among institutions for medical imaging. *J. Am. Med. Inform. Assoc.* 25, 945–954. doi: 10.1093/jamia/ocy017

Cheng, C. L., Azhar, R., Sng, S. H. A., Chua, Y. Q., Hwang, J. S. G., Chin, J. P. F., et al. (2016). Enabling digital pathology in the diagnostic setting: navigating through the implementation journey in an academic medical centre. *J. Clin. Pathol.* 69, 784–792. doi: 10.1136/jclinpath-2015-203600

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

Coroller, T. P., Grossmann, P., Hou, Y., Rios, V., Leijenaar, R. T., Hermann, G., et al. (2016). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* 114, 345–350. doi: 10.1016/j.radonc.2015.02.015

Crimi, A., Bakas, S., Kuijf, H., Menze, B., and Reyes, M. (2018). "Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries," in *Proceedings of the Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017*, Quebec City, QC.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409

Fabelo, H., Ortega, S., Ravi, D., Kiran, B. R., Sosa, C., Bulters, D., et al. (2018). Spatio-spectral classification of hyperspectral images for brain cancer detection during surgical operations. *PLoS One* 13:e0193721. doi: 10.1371/journal.pone.0193721

Fan, R.-E., Chang, W. K., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: a library for large linear classification. *J. Machine Learn. Res.* 9, 1871–1874. doi: 10.1021/ci100073w

Foran, D. J., Yang, L., Chen, W., Hu, J., Goodell, L. A., Reiss, M., et al. (2011). ImageMiner: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. *J. Am. Med. Inform. Assoc.* 18, 403–415. doi: 10.1136/amiajnl-2011-000170

Gao, Y., Ratner, V., Zhu, L., Diprima, T., Kurc, T., Tannenbaum, A., et al. (2016). "Hierarchical nucleus segmentation in digital pathology images," in *Proceedings of the SPIE International Society Optical Engineering*, Washington, DC.

Gillies, R. (2013). Radiomics: informing cancer heterogeneity. *J. Nucl Med.* 31, 271–279.

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: images are more than pictures, they are data. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Griffin, J., and Treanor, D. (2017). Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology* 70, 134–145. doi: 10.1111/his.12993

Gurcan, M. N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., and Yener, B. (2009). Histopathological image analysis: a review. *IEEE Rev. Biomed. Eng.* 2:147. doi: 10.1109/RBME.2009.2034865

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-cnn," in *Proceedings of the IEEE International Conference On Computer Vision*, Seattle, 2961–2969.

Hou, L., Agarwal, A., Samaras, D., Kurc, T. M., Gupta, R. R., and Saltz, J. H. (2019a). "Robust histopathology image analysis: to label or to synthesize?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 8533–8542.

Hou, L., Nguyen, V., Kanevsky, A. B., Samaras, D., Kurc, T. M., Zhao, T., et al. (2019b). Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern Recogn.* 86, 188–200. doi: 10.1016/j.patcog.2018.09.007

Hu, Z. L., Tang, J. S., Wang, Z. M., Zhang, K., Zhang, L., and Sun, Q. L. (2018). Deep learning for image-based cancer detection and diagnosis - A survey. *Pattern Recogn.* 83, 134–149. doi: 10.1016/j.patcog.2018.05.014

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 4700–4708.

Ishikawa, Y., Washiya, K., Aoki, K., and Nagahashi, H. (2016). "Brain tumor classification of microscopy images using deep residual learning," in *SPIE Biophotonics Australasia*, eds M. R. Hutchinson, and E. M. Goldys, (Washington, DC: SPIE), 10013.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull. Soc. Vaudoise Sci. Nat.* 37, 241–272.

Kelahan, L. C., Kalaria, A. D., and Filice, R. W. (2017). PathBot: a radiology-pathology correlation dashboard. *J. Digit. Imaging* 30, 681–686. doi: 10.1007/s10278-017-9969-2

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., and Hajirasouliha, I. (2018). Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBiomedicine* 27, 317–328. doi: 10.1016/j.ebiom.2017.12.026

Kong, J., Cooper, L. A. D., Wang, F., Gutman, D. A., Gao, J., Chisolm, C., et al. (2011). Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes. *IEEE Trans. Biomed. Eng.* 58, 3469–3474. doi: 10.1109/TBME.2011.2169256

Kothari, S., Phan, J. H., Osunkoya, A. O., and Wang, M. D. (2012). Biological interpretation of morphological patterns in histopathological whole-slide images. *ACM BCB* 2012, 218–225. doi: 10.1145/2382936.2382964

Kothari, S., Phan, J. H., Stokes, T. H., and Wang, M. D. (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *J. Am. Med. Inform. Assoc.* 20, 1099–1108. doi: 10.1136/amiajnl-2012-001540

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, eds M. I. Jordan, Y. LeCun, and S. A. Solla, (Cambridge, MA: MIT Press), 1097–1105.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., et al. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* 48, 441–446. doi: 10.1016/j.ejca.2011.11.036

Lee, J. J., Jedrych, J., Pantanowitz, L., and Ho, J. (2018). Validation of digital pathology for primary histopathological diagnosis of routine, inflammatory dermatopathology cases. *Am. J. Dermatopathol.* 40, 17–23. doi: 10.1097/dad.0000000000000888

Lehrer, M., Powell, R. T., Barua, S., Kim, D., Narang, S., and Rao, A. (2017). "Radiogenomics and histomics in glioblastoma: the promise of linking image-derived phenotype with genomic information," in *Advances in Biology and Treatment of Glioblastoma*, ed. K. Somasundaram, (Berlin: Springer), 143–159. doi: 10.1007/978-3-319-56820-1_6

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, (Berlin: Springer), 740–755. doi: 10.1007/978-3-319-10602-1_48

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). "Isolation forest," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, (Amsterdem: IEEE), 413–422.

Lu, S. Y., Lu, Z. H., and Zhang, Y. D. (2019). Pathological brain detection based on AlexNet and transfer learning. *J. Comput. Sci.* 30, 41–47. doi: 10.1016/j.jocs.2018.11.008

Lundstrom, C. F., Gilmore, H. L., and Ros, P. R. (2017). Integrated diagnostics: the computational revolution catalyzing cross-disciplinary practices in radiology. *Pathol. Genomics Radiol.* 285, 12–15. doi: 10.1148/radiol.2017170062

Madabhushi, A., and Lee, G. (2016). Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal.* 33, 170–175. doi: 10.1016/j.media.2016.06.037

Mahmood, F., Borders, D., Chen, R., McKay, G. N., Salimian, K. J., Baras, A., et al. (2018). Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans. Med. Imaging* 36, 1–10. doi: 10.1109/TMI.2019.2927182

Manivannan, S., Li, W., Akbar, S., Wang, R., Zhang, J., and McKenna, S. J. (2016). An automated pattern recognition system for classifying indirect immunofluorescence images of HEp-2 cells and specimens. *Pattern Recogn.* 51, 12–26. doi: 10.1016/j.patcog.2015.09.015

McGarry, S. D., Hurrell, S. L., Iczkowski, K. A., Hall, W., Kaczmarowski, A. L., Banerjee, A., et al. (2018). Radio-pathomic maps of epithelium and lumen density predict the location of high-grade prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 101, 1179–1187. doi: 10.1016/j.ijrobp.2018.04.044

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., et al. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U.S.A.* 115, E2970–E2979. doi: 10.1073/pnas.1717139115

Momeni, A., Thibault, M., and Gevaert, O. (2018). "Dropout-enabled ensemble learning for multi-scale biomedical data," in *Proceedings of the International MICCAI Brainlesion Workshop*, (Berlin: Springer), 407–415. doi: 10.1007/978-3-030-11723-8_41

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., and Aerts, H. J. (2015). Machine learning methods for quantitative radiomic biomarkers. *Sci. Rep.* 5:13087. doi: 10.1038/srep13087

Peikari, M., and Martel, A. L. (2016). "Automatic cell detection and segmentation from H and E stained pathology slides using colorspace decorrelation stretching," in *Proceedings of the Medical Imaging 2016: Digital Pathology, International Society for Optics and Photonics*, Vol. 9791, Houston, TX, 979114.

Peikari, M., Salama, S., Nofech-Mozes, S., and Martel, A. L. (2018). A cluster-then-label semi-supervised learning approach for pathology image classification. *Sci. Rep.* 8:7193. doi: 10.1038/s41598-018-24876-0

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Amsterdam: Elsevier.

Prior, F. W., Clark, K., Commean, P., Freymann, J., Jaffe, C., Kirby, J., et al. (2013). "TCIA: an information resource to enable open science," in *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine (EMBC)*, (Piscataway, NJ: IEEE), 1282–1285.

Qi, Q., Li, Y., Wang, J., Zheng, H., Huang, Y., Ding, X., et al. (2018). Label-efficient breast cancer histopathological image classification. *IEEE J. Biomed. Health Inform.* 23, 2108–2116. doi: 10.1109/JBHI.2018.2885134

Qian, Z., Li, Y., Wang, Y., Li, L., Li, R., Wang, K., et al. (2019). Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers. *Cancer Lett.* 451, 128–135. doi: 10.1016/j.canlet.2019.02.054

Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *IEEE Comput. Graph. Appl.* 21, 34–41.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster r-cnn: towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, eds M. I. Jordan, Y. LeCun, and S. A. Solla, (Cambridge, MA: MIT Press), 91–99.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Berlin: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Saltz, J., Almeida, J., Gao, Y., Sharma, A., Bremer, E., DiPrima, T., et al. (2017). Towards generation, management, and exploration of combined radiomics and pathomics datasets for cancer research. *AMIA Jt Summits Transl. Sci. Proc.* 2017, 85–94.

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., et al. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* 23:181-193.e7. doi: 10.1016/j.celrep.2018.03.086

Senaras, C., Niazi, M. K. K., Lozanski, G., and Gurcan, M. N. (2018a). DeepFocus: detection of out-of-focus regions in whole slide digital images using deep learning. *PLoS One* 13:e0205387. doi: 10.1371/journal.pone.0205387

Senaras, C., Niazi, M. K. K., Sahiner, B., Pennell, M. P., Tozbikian, G., Lozanski, G., et al. (2018b). Optimized generation of high-resolution phantom images using cGAN: application to quantification of Ki67 breast cancer images. *PLoS One* 13:e0196846. doi: 10.1371/journal.pone.0196846

Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., and Bakas, S. (2018). "Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation," in *Proceedings of the International MICCAI Brainlesion Workshop*, (Berlin: Springer), 92–104. doi: 10.1007/978-3-030-11723-8_9

Shukla, G., Alexander, G. S., Bakas, S., Nikam, R., Talekar, K., Palmer, J. D., et al. (2017). Advanced magnetic resonance imaging in glioblastoma: a review. *Chin. Clin. Oncol.* 6:40. doi: 10.21037/cco.2017.06.28

Siegel, R. L., Miller, K. D., and Jemal, A. (2016). Cancer statistics, 2016. *CA* 66, 7–30. doi: 10.3322/caac.21332

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA* 69, 7–34.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [preprint]

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19:A68. doi: 10.5114/wo.2014.47136

van Griethuysen, J., Fedorov, A., Aucoin, N., Fillion-Robin, J.-C., Hosny, A., Pieper, S., et al. (2016). *Welcome to Pyradiomics Documentation*. Available at: https://pyradiomics.readthedocs.io/en/latest/ (accessed February 7, 2020).

van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Vu, Q. D., Graham, S., Kurc, T., To, M. N. N., Shaban, M., Qaiser, T., et al. (2019). Methods for segmentation and classification of digital microscopy tissue images. *Front. Bioeng. Biotechnol.* 7:53. doi: 10.3389/fbioe.2019.00053

Wang, X. Y., Wang, D. Q., Yao, Z. G., Xin, B. W., Wang, B., Lan, C. J., et al. (2019). Machine learning models for multiparametric glioma grading with quantitative result interpretations. *Front. Neurosci.* 12:1046. doi: 10.3389/fnins.2018.01046

Wen, S., Kurc, T. M., Gao, Y., Zhao, T., Saltz, J. H., and Zhu, W. (2017). A methodology for texture feature-based quality assessment in nucleus segmentation of histopathology image. *J. Pathol. Inform.* 8:38. doi: 10.4103/jpi.jpi_43_17

Wen, S., Kurc, T. M., Hou, L., Saltz, J. H., Gupta, R. R., Batiste, R., et al. (2018). Comparison of different classifiers with active learning to support quality control in nucleus segmentation in pathology images. *AMIA Jt Summits Transl. Sci. Proc.* 2017, 227–236.

Wollmann, T., Bernhard, P., Gunkel, M., Braun, D. M., Meiners, J., Simon, R., et al. (2019). "Black-box hyperparameter optimization for nuclei segmentation in prostate tissue images," in *Bildverarbeitung für die Medizin 2019*, eds H.

Handels, T. M. Deserno, A. Maier, K. H. Maier-Hein, and C. Palm, (Berlin: Springer), 345–350. doi: 10.1007/978-3-658-25326-4_75

Xie, L., and Li, C. (2018). "Simultaneous detection and segmentation of cell nuclei based on convolutional neural network," in *Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine*, (Chengdu: ACM), 129–132.

Xing, F., and Yang, L. (2016). Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE Rev. Biomed. Eng.* 9, 234–263. doi: 10.1109/RBME.2016.2515127

Yang, Q., Wu, K., Cheng, H., Gu, C., Liu, Y., Casey, S. P., et al. (2018). "Cervical nuclei segmentation in whole slide histopathology images using convolution neural network," in *Proceedings of the International Conference on Soft Computing in Data Science*, (Berlin: Springer), 99–109. doi: 10.1007/978-981-13-3441-2_8

Yonekura, A., Kawanaka, H., Prasath, V. B. S., Aronow, B. J., and Tsuruoka, S. (2018). "Glioma subtypes clustering method using histopathological image analysis," in *Proceedings of the 7th International Conference on Informatics, Electronics & Vision (Iciev) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (Icivpr), 2018*, Fukuoka, 442–446.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., et al. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* 7:12474. doi: 10.1038/ncomms12474

Yuan, Y., Shi, Q., Li, M., Nagamuthu, C., Andres, E., and Davis, F. G. (2016). Canadian brain cancer survival rates by tumour type and region: 1992–2008. *Can. J. Public Health* 107, e37–e42. doi: 10.17269/cjph.107.5209

Zhou, M., Scott, J., Chaudhury, B., Hall, L., Goldgof, D., Yeom, K. W., et al. (2018). Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *Am. J. Neuroradiol.* 39, 208–216. doi: 10.3174/ajnr.A5391

Zwanenburg, A., Leger, S., Vallières, M., and Löck, S. (2016). Image biomarker standardisation initiative. *arXiv [preprint]*

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 20 March 2020
doi: 10.3389/fncom.2020.00017

Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation

Théo Estienne[1,2,3,4*†], *Marvin Lerousseau*[1,2,3,5†], *Maria Vakalopoulou*[1,4,5], *Emilie Alvarez Andres*[1,2,3], *Enzo Battistella*[1,2,3,4], *Alexandre Carré*[1,2,3], *Siddhartha Chandra*[5], *Stergios Christodoulidis*[6], *Mihir Sahasrabudhe*[5], *Roger Sun*[1,2,3,5], *Charlotte Robert*[1,2,3], *Hugues Talbot*[5], *Nikos Paragios*[1] *and Eric Deutsch*[1,2,3]

[1] *Gustave Roussy-CentraleSupélec-TheraPanacea Center of Artificial Intelligence in Radiation Therapy and Oncology, Gustave Roussy Cancer Campus, Villejuif, France,* [2] *Université Paris-Saclay, Institut Gustave Roussy, Inserm, Molecular Radiotherapy and Innovative Therapeutics, Villejuif, France,* [3] *Gustave Roussy Cancer Campus, Department of Radiation Oncology, Villejuif, France,* [4] *Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, Gif-sur-Yvette, France,* [5] *Université Paris-Saclay, CentraleSupélec, Inria, Centre de Vision Numérique, Gif-sur-Yvette, France,* [6] *Université Paris-Saclay, Institut Gustave Roussy, Inserm, Predictive Biomarkers and Novel Therapeutic Strategies in Oncology, Villejuif, France*

Image registration and segmentation are the two most studied problems in medical image analysis. Deep learning algorithms have recently gained a lot of attention due to their success and state-of-the-art results in variety of problems and communities. In this paper, we propose a novel, efficient, and multi-task algorithm that addresses the problems of image registration and brain tumor segmentation jointly. Our method exploits the dependencies between these tasks through a natural coupling of their interdependencies during inference. In particular, the similarity constraints are relaxed within the tumor regions using an efficient and relatively simple formulation. We evaluated the performance of our formulation both quantitatively and qualitatively for registration and segmentation problems on two publicly available datasets (BraTS 2018 and OASIS 3), reporting competitive results with other recent state-of-the-art methods. Moreover, our proposed framework reports significant amelioration ($p < 0.005$) for the registration performance inside the tumor locations, providing a generic method that does not need any predefined conditions (e.g., absence of abnormalities) about the volumes to be registered. Our implementation is publicly available online at https://github.com/TheoEst/joint_registration_tumor_segmentation.

Keywords: brain tumor segmentation, deformable registration, multi-task networks, deep learning, convolutional neural networks

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Artur Luczak,
University of Lethbridge, Canada
Guotai Wang,
University of Electronic Science and Technology of China, China

****Correspondence:***
Théo Estienne
theo.estienne@centralesupelec.fr

† These authors have contributed equally to this work

Received: *03 July 2019*
Accepted: *11 February 2020*
Published: *20 March 2020*

Citation:
Estienne T, Lerousseau M, Vakalopoulou M, Alvarez Andres E, Battistella E, Carré A, Chandra S, Christodoulidis S, Sahasrabudhe M, Sun R, Robert C, Talbot H, Paragios N and Deutsch E (2020) Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation. Front. Comput. Neurosci. 14:17. doi: 10.3389/fncom.2020.00017

1. INTRODUCTION

Brain tumors and more specifically gliomas as one of the most frequent types, are across the most dangerous and rapidly growing types of cancer (Holland, 2002). In clinical practice, multi-modal magnetic resonance imaging (MRI) is the primary method of screening and diagnosis of gliomas. While gliomas are commonly stratified into Low grade and High grade due to different histology and imaging aspects, prognosis and treatment strategy, radiotherapy is one of the mainstays of treatment (Stupp et al., 2014; Sepúlveda-Sánchez et al., 2018).

However, radiotherapy treatment planning relies on tumor manual segmentation by physicians, making the process tedious, time-consuming, and sensitive to bias due to low inter-observer agreement (Wee et al., 2015).

In order to overcome these limitations, numerous methods have been proposed recently that try to provide tools and algorithms that will make the process of gliomas segmentation automatic and accurate (Parisot et al., 2016; Zhao et al., 2018). Toward this direction, the multimodal brain tumor segmentation challenge (BraTS) (Menze et al., 2015; Bakas et al., 2017a,b,c) is annually organized, in order to highlight efficient approaches and indicate the way toward this challenging problem. In recent years, most of the approaches that exploit BraTS have been based on deep learning architectures using 3D convolutional neural networks (CNNs) similar to VNet (Milletari et al., 2016). In particular, the best performing approaches use ensembles of deep learning architectures (Kamnitsas et al., 2018; Zhou et al., 2018), with autoencoder regularization (Myronenko, 2018) or they even combine deep learning architectures together with algorithms, such as conditional random fields (CRFs) (Chandra et al., 2019). Other top-performing methods in the BraTS 2017 and 2018 challenges used cascaded networks, multi-view and multi-scale approaches (Wang et al., 2017), generic UNet architecture with data augmentation and post-processing (Isensee et al., 2018), dilated convolutions and label uncertainty loss (McKinley et al., 2018), and context aggregation and localization pathways (Isensee et al., 2017). A more detailed comparison and presentation of competing methods in recent BraTS challenges is presented and summarized in Bakas et al. (2018).

Image registration is a challenging task for medical image analysis in general and for rapidly evolving brain tumors in particular, where longitudinal assessment is critical. Image registration seeks to determine a transformation that will map two volumes (source and reference) to the same coordinate system. In practice, we seek a volume mapping function that changes the coordinate system of the source volume into the coordinate system of the reference volume. Among the different types of methods employed in medical applications, deformable or elastic registration is the most commonly used (Sotiras et al., 2013). Linear methods are an alternative but in that case a linear global transformation is sought for the entire volume. Deformable registration has been addressed with a variety of methods, including for example surface matching (Postelnicu et al., 2009; Robinson et al., 2018) or graph based approaches (Glocker et al., 2009). These methods have been extended to address co-registration of multiple volumes (Ou et al., 2011). Moreover, some of the most popular methods traditionally used for the accurate deformable registration include (Avants et al., 2008; Klein et al., 2009; Shi et al., 2013). Recently a variety of deep learning based methods have been proposed, reducing significantly the computational time but maintaining the accuracy and robustness of the registration (Christodoulidis et al., 2018; Dalca et al., 2018). In particular, the authors in Dalca et al. (2018) presented a deep learning framework trained for atlas-based registration of brain MR images, while in Christodoulidis et al. (2018) the authors present a scheme for a concurrent linear and deformable registration of lung MR images. However, when it comes to anatomies that contain abnormalities, such as tumoral areas, these methods fail to register the volumes at certain locations, due to lack of similarity between them. This often leads to distortions in and around the tumor regions in the deformed image.

To overcome this problem, in this paper, we propose a dual deep learning based architecture that addresses registration and tumor segmentation simultaneously, relaxing the registration constraints inside the predicted tumor areas, providing displacements and segmentation maps at the same time. Our framework bears concept similarities with the work presented in Parisot et al. (2012) where a Markov Random Field (MRF) framework has been proposed to address both of tumor segmentation and image registration jointly. Their method required ∼6 min for the registration of one pair and the segmentation of one class tumor region was performed with handcrafted features and classical machine learning techniques using only one MRI modality. Moreover, there are methods in the literature that try to address the problem of registration of brain tumor MRI by registering on atlases or MRIs without tumoral regions (Gooya et al., 2010, 2012). Here, we introduce a highly scalable, modular, generic, and precise 3D-CNN for both registration and segmentation tasks and provide a computationally efficient and accurate method for registering any arbitrary subject involving possible abnormalities. To the best of our knowledge this is the first time that a joint deep learning-based architecture is presented, showing very promising results in two publicly available datasets for brain MRI. The proposed framework provides a very powerful formulation by introducing the means to elucidate clinical or functional trends in the anatomy or physiology of the brain via the registration branch. It further enables the modeling and the detection of brain tumor areas due to the synergy with the segmentation branch.

2. MATERIALS AND METHODS

Consider a pair of medical volumes from two different patients—a source S, and a reference R together with their annotations for the tumor areas (S_{seg} and R_{seg}). The framework consists of a bi-cephalic structure with shared parameters, depicted in **Figure 1**. During training the network uses as input a source S and a reference R volumes and outputs their brain tumor segmentation masks $\widehat{S}_{seg}$ and $\widehat{R}_{seg}$ and the optimal elastic transformation G which will project or map the source volume to the reference volume. The goal of the registration part is to find the optimal transformation to transform the source S to the reference R volume. In this section, we present the details for each of the blocks as well as our final formulation for the optimization.

2.1. Shared Encoder

One of the main differences of the proposed formulation with other registration approaches in the literature is the way that the source and reference volumes are combined. In particular, instead of concatenating the two initial volumes, these volumes are independently forwarded in a unique encoder, yielding two sets of features maps (called *latent codes*) C_{source} and $C_{\text{reference}}$ for the source and the reference volumes, respectively. These two codes are then independently forwarded into the segmentation decoder, providing the predicted segmentation maps $\widehat{S}_{seg}$ and

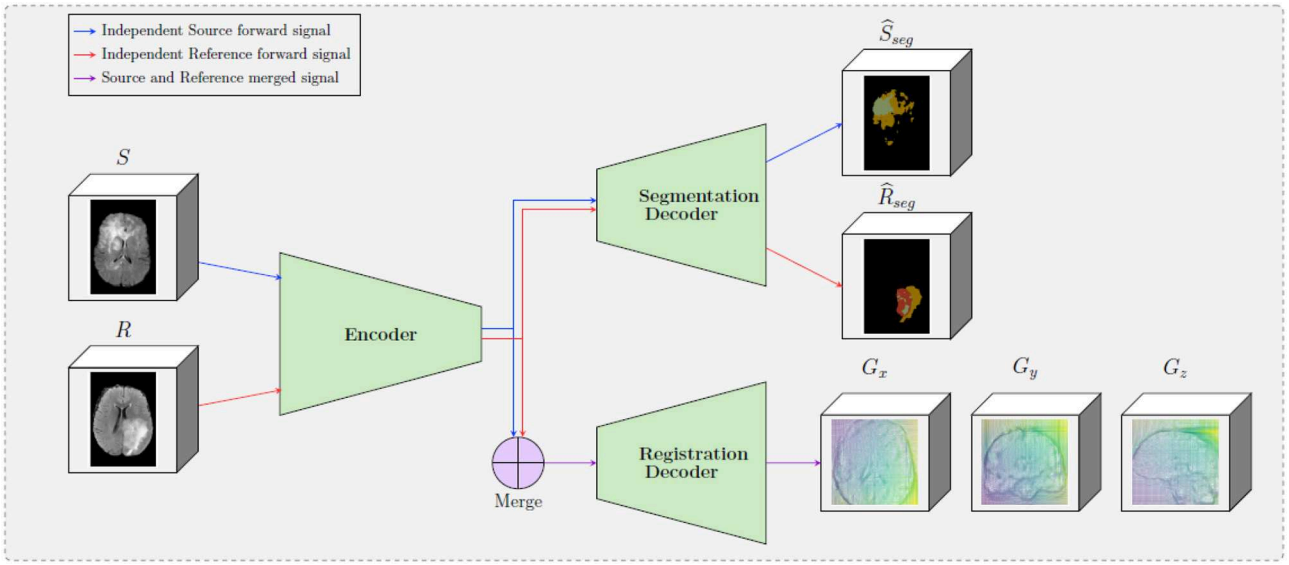


FIGURE 1 | A schematic representation of the proposed framework. The framework is composed by two decoders, one which provides tumor segmentation masks for both *S* and *R* images, and one the provides the optimal displacement grid *G* that will accurately map the *S* to the *R* image. The merge bloc will combine the forward signal of the source input and the reference input (which are forwarded independently in the encoder).

$\widehat{R}_{seg}$. Simultaneously, the two codes are merged before being forwarded in the registration decoder—this operation is depicted in the "Merge" block in **Figure 1**. The motivation behind adopting this strategy is based on forcing the encoder to extract meaningful representations from individual volumes instead of a pair of volumes. This is equivalent to asking the encoder discovering a template, "deformation-free" space for all volumes, and encoding each volume against this space (Shu et al., 2018), instead of decoding the deformation grid between every possible pair of volumes. Besides, from the segmentation point of view, there are no relationship between the tumor maps of the source volume and the reference volume, so the codes to be forwarded into the segmentation decoder should not depend on each other.

We tested two merging operators, namely concatenation and subtraction. Both source and reference images are $4D$ volumes whose first dimension corresponds to the 4 different MRI modalities that are used per subject. After the forward to the encoder, the codes C_{source} and $C_{\text{reference}}$ are also 4D volumes with the first dimension corresponding to n_f, which is the number of convolutional filters of the last block of the encoder. Before C_{source} and $C_{\text{reference}}$ are inserted into the registration decoder, they are merged, outputting one 4D volume of size $2 \times n_f$ in the case of the concatenation, and of size n_f for the elementwise subtraction operator, both leaving the rest of the dimensions unchanged. In particular, the subtraction presents the following natural properties for every coding image C_I:

- $\forall C_I \in \mathbb{R}^n : Merge(C_I, C_I) = 0$
- $\forall C_I, C_J \in \mathbb{R}^n \times \mathbb{R}^n : Merge(C_I, C_J) = -Merge(C_J, C_I)$

2.2. Brain Tumor Segmentation Decoder

Inspired by the latest advances reported on the BraTS 2018 dataset, we adopt a powerful autoencoder architecture. The segmentation and registration decoders share the same encoder (section 2.1) for feature extraction and they provide brain tumor segmentation masks ($\widehat{S}_{seg}$ and $\widehat{R}_{seg}$) for the source and the reference images. These masks refer to valuable information about the regions that cannot be registered properly as there is no corresponding anatomical information on the pair. This information is integrated into the optimization of the registration component, relaxing the similarity constraints and preserving to a certain extent the geometric properties of the tumor.

Variety of loss functions have been proposed in the literature for the semantic segmentation of 3D medical volumes. In this paper, we performed all our experiments using weighted categorical cross-entropy loss and optimizing three different segmentation classes for the tumor area as provided by the BraTS dataset. In particular,

$$\mathcal{L}_{seg} = CE(S_{seg}, \widehat{S}_{seg}) + CE(R_{seg}, \widehat{R}_{seg}) \tag{1}$$

where CE denotes the weighted cross entropy loss. The cross entropy is calculated for both the source and reference images and the overall segmentation loss is the sum of the two. Here we should note that different segmentation losses can be applicable as for example the dice coefficient (Sudre et al., 2017), focal loss (Lin et al., 2017), etc.

2.3. Elastic Registration Decoder

In this paper, the registration strategy is based on the one presented in Christodoulidis et al. (2018), with the main component being the 3D spatial transformer. A spatial transformer deforms (or warps) a given image S with a deformation grid G. It can be represented by the operation,

$$D = \mathcal{W}(S, G),$$

where $\mathcal{W}(\cdot, G)$ indicates a sampling operation $\mathcal{W}$ under the deformation G and D the deformed image. The deformation is hence fed to the transformer layer as sampling coordinates for a backward trilinear interpolation sampling, adapting a strategy similar to Shu et al. (2018). The sampling process is then described by

$$D(\vec{p}) = \mathcal{W}(S, G)(\vec{p}) = \sum_{\vec{q}} S(\vec{q}) \prod_{d} \max\left(0, 1 - \left|[G(\vec{p})]_d - \vec{q}_d\right|\right),$$

where $\vec{p}$ and $\vec{q}$ denote pixel locations, $d \in \{x, y, z\}$ denotes an axis, and $[G(\vec{p})]_d$ denotes the d-component of $G(\vec{p})$. Moreover, instead of regressing per-pixel displacements, we predict a matrix Ψ of spatial gradients between consecutive pixels along each axis. The actual grid G can then be obtained by applying an integration operation on Ψ along the x-, y-, and z-axes, which is approximated by the cumulative sum in the discrete case. Consequently, two pixels $\vec{p}$ and $\vec{p} + 1$ will have moved closer, maintained distance, or moved apart in the warped image, if $\Psi_{\vec{p}}$ is respectively < 1, $= 1$, or > 1.

2.4. Network Architecture

Our network architecture is a modified version of the fully convolutional VNet (Milletari et al., 2016) for the underlying encoder and decoders parts, maintaining the depth of the model and the rest of the filter's configuration unchanged. The model, whose computational graph is displayed in **Table 1**, comprises several sequential residual convolutional blocks made of one to three convolutional layers, followed by downsampling convolutions for the encoder part and upsampling convolutions for the decoder part. We replaced the initial $5 \times 5 \times 5$ convolutions filter-size by $3 \times 3 \times 3$ in order to reduce the number of parameters without changing the depth of the model, and also replace PReLu activations by ReLU ones. In order to speed up its convergence, the model uses residual connections between each encoding and corresponding decoding stage for both the segmentation and the registration decoder. This allows every layer of the network, particularly the first ones, to be trained more efficiently since the gradient can flow easier from the last layers to the first ones with less vanishing or exploding gradient issues. The encoder part deals with 4-inputs per volume, representing the four different MRI modalities that are available on the BraTS dataset, an extra $1 \times 1 \times 1$ convolution is added to fuse the initial modalities. Moreover, the architecture contains 2 decoders of identical blocks, 1 dedicated to the segmentation of tumors for the source and reference image and 1 dedicated to the optimal displacement that will map the source to the reference image.

2.5. Optimization

The network is trained to minimize the segmentation and registration loss functions jointly. For the segmentation task the loss function is summarized in Equation (1). For registration, the classical optimization scheme is to minimize the Frobenius norm between the R and D image intensities:

$$\mathcal{L}_{reg} = ||(R - D)||^2 + \alpha \left\|\Psi - \Psi_I\right\|_1 \tag{2}$$

Here, in order to better achieve overall registration, the Frobenius norm within the regions predicted to be tumors is excluded from the loss function. We argue that by doing this, the model does not focus on tumor regions, which might produce very high norm due to their texture, but rather focuses on the overall registration task by looking at regions outside the tumor which contain information more pertinent to the alignment of the volumes. Here we should mention that on $\widehat{S}_{seg}$ we apply the same displacement grid as on S, resulting in $D_{seg} = \mathcal{W}(\widehat{S}_{seg}, G)$. Further, let $\widehat{R}^0_{seg}$ and D^0_{seg} be binary volumes indicating the voxels which are predicted to be outside any segmented regions. Then, the registration loss can be written as

$$\mathcal{L}^{\star}_{reg} = ||(R - D) \cdot D^0_{seg} \cdot \widehat{R}^0_{seg}||^2 + \alpha \left\|\Psi - \Psi_I\right\|_1 \tag{3}$$

where $\cdot$ is the element-wise multiplication, $||\cdot||^2$ indicates the Frobenius norm, Ψ_I is the spatial gradient of the identity deformation and α is the regularization hyperparameter. The use of regularization on the displacements Ψ is essential in order to constrain the network to predict smooth deformation grids that are anatomically more meaningful while at the same time regularize the objective function toward avoiding local minimum.

Finally the final optimization of the framework is performed by the joint optimization of the segmentation and registration loss functions

$$\mathcal{L} = \mathcal{L}_{reg} + \beta \mathcal{L}_{seg}$$

where β is a weight that indicates the influence of each of the components on the joint optimization of the network and was defined after grid search.

For the training process, the initial learning rate was $2 \cdot 10^{-3}$ and subdued by a factor of 5 if the performance on the validation set did not improve for 30 epochs. The training procedure stops when there is no improvement for 50 epochs. The regularization weights α and β were set to 10^{-10} and 1 after grid search. As training samples, random pairs among all cases were selected with a batch size limited to 2 due to the limited memory resources on the GPU. The performance of the network was evaluated every 100 batches, and both proposed models converged after nearly 200 epochs. The overall training time was calculated to ∼20 h, while the time for inference of one pair, using four different modalities was ∼3 s, using an NVIDIA GeForce GTX 1080 Ti GPU.

2.6. Datasets

We evaluated the performance of our method using two publicly available datasets, namely the Brain Tumor Segmentation (BraTS) (Bakas et al., 2018) and Open Access Series of Imaging Studies (OASIS 3) (Marcus et al., 2010) datasets. BraTS contains multi-institutional pre-operative MRI scans of whole brains with visible gliomas, which are intrinsically heterogeneous in their imaging phenotype (shape and appearance) and histology. The MRIs are all pre-operative and consist of four modalities, i.e., 4 3D volumes, namely (a) a native T1-weighted scan (T1),

TABLE 1 | Layer architectures of the shared encoder, the segmentation decoder and the registration decoder.

Name	Input	Res. input	Operations	Output shape
ENCODER				
Enc^1	4D MRI		$Conv_{1,8}$, ReLU, ($Conv_{3,8}$, ReLU), AddId,	(144, 208, 144, 8)
Enc^2	Enc^1		$Conv_{2,16}$, ReLU, ($Conv_{3,16}$, ReLU)*2, AddId	(72, 104, 72, 16)
Enc^3	Enc^2		$Conv_{2,32}$, ReLU, ($Conv_{3,32}$, ReLU)*3, AddId	(36, 52, 36, 32)
Enc^4	Enc^3		$Conv_{2,64}$, ReLU, ($Conv_{3,64}$, ReLU)*3, AddId	(18, 26, 18, 64)
Enc^5	Enc^4		$Conv_{2,128}$, ReLU, ($Conv_{3,128}$, ReLU)*3, AddId	(9, 13, 9, 128)
SEGMENTATION DECODER				
Dec^4_{seg}	Enc^5	Enc^4	$DeConv_{2,64}$,ReLU, ResConc, ($Conv_{3,64}$, ReLU)*3, AddId	(18, 26, 18, 64)
Dec^3_{seg}	Dec^4_{seg}	Enc^3	$DeConv_{2,32}$, ReLU, ResConc, ($Conv_{3,32}$, ReLU)*3, AddId	(36, 52, 36, 32)
Dec^2_{seg}	Dec^3_{seg}	Enc^2	$DeConv_{2,16}$, ReLU, ResConc, ($Conv_{3,16}$, ReLU)*2, AddId	(72, 104, 72, 16)
Dec^1_{seg}	Dec^2_{seg}	Enc^1	$DeConv_{2,8}$, ReLU, ResConc, ($Conv_{3,8}$, ReLU), AddId	(144, 208, 144, 8)
Dec^0_{seg}	Dec^1_{seg}		$Conv_{1,4}$, Softmax	(144, 208, 144, 4)
REGISTRATION DECODER				
Merge	Enc_R^j, Enc_S^j		For all $1 \leq i \leq 5$, $\mathrm{MEnc}^i = \mathrm{Enc}_R^i \oplus \mathrm{Enc}_S^i$	
Dec^4_{reg}	$MEnc^5$	$MEnc^4$	$DeConv_{2,64}$, ReLU, ResConc, ($Conv_{3,64}$, ReLU)*3, AddId	(18, 26, 18, 64)
Dec^3_{reg}	Dec^4_{reg}	$MEnc^3$	$DeConv_{2,32}$, ReLU, ResConc, ($Conv_{3,32}$, ReLU)*3, AddId	(36, 52, 36, 32)
Dec^2_{reg}	Dec^3_{reg}	$MEnc^2$	$DeConv_{2,16}$, ReLU, ResConc, ($Conv_{3,16}$, ReLU)*2, AddId	(72, 104, 72, 16)
Dec^1_{reg}	Dec^2_{reg}	$MEnc^1$	$DeConv_{2,8}$, ReLU, ResConc, ($Conv_{3,8}$, ReLU), AddId	(144, 208, 144, 8)
Dec^0_{reg}	Dec^1_{reg}		$Conv_{1,3}$, Sigmoid	(144, 208, 144, 3)

*The sub-architectures are grouped into blocks, one per table line, whose names are indicated in the first column. Each block processed a forward signal as input identified by the second column. Additionally, both decoders have residual connections from different stages of the encoder, identified by the third column. The blocks are made of a set of successive operations where $Conv_{w,f}$ (resp. $DeConv_{w,f}$) stands for a convolutional (resp. deconvolutional) layer with weight size $w \times w \times w$ and f filters, ReLU—Rectified Linear Unit, AddId—intra-block residual connection with the output of the first activated convolution of the corresponding block, ResConc—encoder to decoder residual connection from the output of the third column block to the current signal, Softmax and Sigmoid—finale output activation. * indicates successive repetition of the previous operations in parenthesis. For convolutions and deconvolutions layers, strides is $1 \times 1 \times 1$ except for the $Conv_{2,}$. which is $2 \times 2 \times 2$. The first layer of the registration decoder indicates the merging operation of the source signal and the reference signal, which are obtained by inferring them successively in the encoder; $\oplus$ Indicates elementwise subtraction or channelwise concatenation of the source and reference list of tensors (forward network signal and four residual connection signals). The last column indicates each block output shape (channels last).*

(b) a post-contrast Gadolinium T1-weighted scan (T1Gd), (c) a native T2-weighted scan (T2), and (d) a native T2 Fluid Attenuated Inversion Recovery scan (T2-FLAIR). The BraTS MRIs are provided with voxelwise ground-truth annotations for five disjoint classes denoting (a) the background, (b) the necrotic and non-enhancing tumor core (NCR/NET), (c) the GD-enhancing tumor (ET), (d) the peritumoral edema (ED) as well as invaded tissue, and finally (e) the rest of the brain, i.e., brain with no abnormality nor invaded tissue. Each center was responsible for annotating their MRIs, with a central validation by domain experts. We use the original dataset split of BraTS 2018 which contains 285 training samples and 66 for validation. In order to perform our experiments, we split this training set into three parts, i.e., train, validation and test sets (199, 26, and 60 patients, respectively), while we used the 66 unseen cases on the platform to report the performance of the proposed and the benchmarked methods. Moreover, and especially for the registration task, we evaluated the performance of the models trained on BraTS on the OASIS 3 dataset to test the generalization of the method. We extract from this dataset a subset of 150 subjects which were characterized as either non-demented or with mild cases of Alzheimer's disease (AD) using the Clinical Dementia Rating (CDR). Each scan is made of 3–4 individual T1-weighted MRIs, which has been intended to reduce the signal-to-noise ratio visible with single images. The scans are also provided with annotations for 47 different structures for left and right side of the brain generated with FreeSurfer. Some samples of both datasets can be seen in **Figure 2**.

The same pre-processing steps have been applied for both datasets. MRIs were resampled to voxels of volume 1 mm^3 using trilinear interpolation. Each scan is then centered by automatically translating their barycenter to the center of the volume. Ground-truth masks of training and validation steps were accordingly translated. Each modality of each scan has been standardized, i.e., the values of the voxels of the 3D subscans were of zero mean and of unit variance. This normalization step is done independently for each patient and for each channel in order to equally consider each channel since modalities have voxels values in completely different ranges. Finally, these consequent scans are cropped into (144, 208, 144) sized volumes.

2.7. Statistical Evaluations

Our contributions in this study are three-fold: (i) a multi-task scheme for joint segmentation and registration; (ii) an encoding scheme followed by a fusion scheme in the latent space to aggregate information from the pair of images; and (iii) a loss formulation (Equation 3) that relaxes the registration constraints in the tumoral regions. In this section, we present

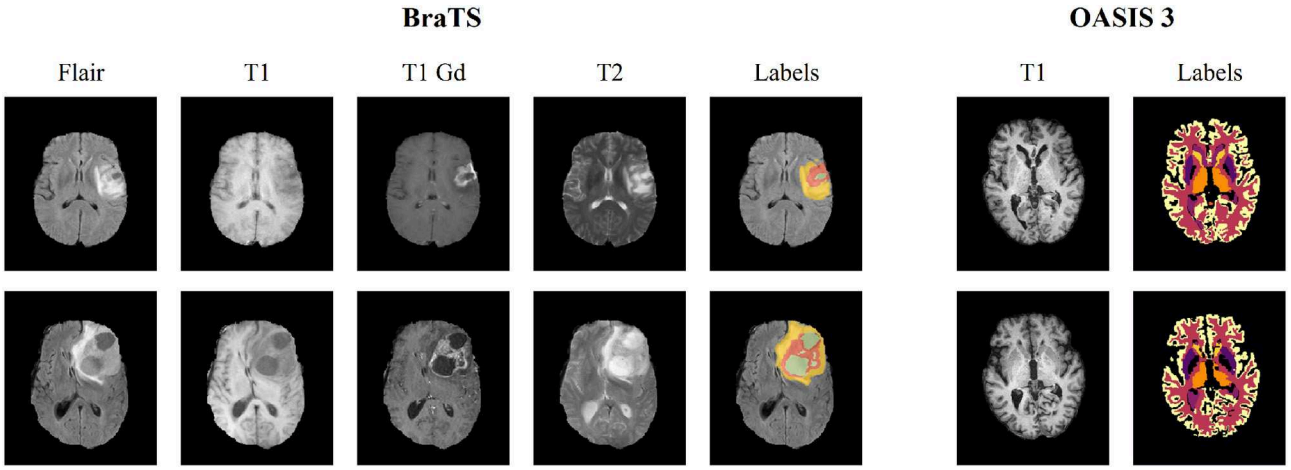


FIGURE 2 | Illustration of a slice extracted from two different subjects for both BraTS 2018 and OASIS 3 datasets. The BraTS dataset consists of four modalities (T1, T1 gadolinium [T1 Gd], T2, T2 FLAIR [Flair]), along with voxelwise annotations for the three tumor tissue subclasses depicting the overall extent of tumors. OASIS 3 consists of a single T1 modality, and images are provided with voxelwise annotations for 47 different normal brain structures for patients without brain tumors.

our extensive experiments to demonstrate the soundness of our method.

2.7.1. Comparison With Competing Methods

To demonstrate the importance of each component of our method, we performed multiple experiments to evaluate performance for both registration and segmentation tasks by removing one or more components. In particular, we evaluated 2 merging operators—subtraction and concatenation. The resulting models are henceforth referred to as "Proposed concatenation with $\mathcal{L}^{\star}_{reg}$" and "Proposed subtraction with $\mathcal{L}^{\star}_{reg}$," respectively. We further evaluated the importance of the proposed loss formulation, reporting the performance of the models without including it in the total loss. This model is called "w/o $\mathcal{L}^{\star}_{reg}$." Finally, we also evaluated the performance of the method without the segmentation decoder, which is reported as "Proposed concatenation only reg." and "Proposed subtraction only reg.," which again did not use $\mathcal{L}^{\star}_{reg}$.

We also benchmark baseline methods, without any of the proposed contributions. Since our deep learning architecture is derived from the Vnet (Milletari et al., 2016), this model is used as baseline for segmentation. This comparison seems fair since the fully proposed approach can be seen as a Vnet for the task of segmentation: the shared encoder and the proposed loss are primarily designed for registration, and have no direct impact on the segmentation apart from the features learnt in the encoder. For completeness, the top performing results on the BraTS (Bakas et al., 2018) challenge are reported, although we argue that the comparison is unfair since our deep learning architecture is entirely based on the Vnet (Milletari et al., 2016), which is not specifically designed to perform well on the BraTS segmentation task. Finally, we also report the performance of Voxelmorph (Dalca et al., 2018), a well-performing brain MRI registration neural network-based approach, although their entire deep learning structure as well as their grid formulation is different.

2.7.2. Performance Assessment

For performance assessment of the segmentation task, we reported the Dice coefficient metric and Hausdorff distance to measure the performance for the tumor classes Tumor Core (TC), Enhancing Tumor (ET), and Whole Tumor (WT) as computed and provided from the BraTS submission website. These classes are the ones used in the BraTS challenge (Bakas et al., 2018), but differ from the original ones provided in the BraTS dataset: TC is the same as the one labeled in the BraTS dataset for necrotic core (NCR/NET), ET is the disjoint union of the original classes NCR/NET and ET, while WT refers to the union of all tumoral and invaded tissues.

For the registration, we evaluated the change on the tumor area together with the Dice coefficient metric for the following categories of the OASIS 3 dataset: brain stem (BS), cerebrospinal fluid (CSF), 4th ventricle (4V), amygdala (Am), caudate (Ca), cerebellum cortex (CblmC), cerebellum white matter (CblmWM), cerebral cortex (CeblC), cerebral white matter (CeblWM), hippocampus (Hi), lateral ventricle (LV), pallidum (Pa), putamen (Pu), ventral DC (VDC), and 3rd ventricle (3V) categories. Here we should mention that for the experiments with the OASIS 3 dataset, we performed a training only with the T1-weighted MRIs of the BraTS dataset, in order to match the available modalities of the OASIS 3 dataset. This evaluation is important as (i) BraTS does not provide anatomical annotations in order to evaluate quantitatively the registration performance and (ii) the generalization of the proposed method on an unseen dataset is evaluated. For the registration of tumor tissues, which might not exist in the source or reference MRIs, we expect the model to register tumor areas while maintaining

their geometric properties. In particular, we do not really expect the tumor areas to stay completely unchanged. However, we expect that the volume of the different tumor types would change with a ratio similar to the one that the entire source to the reference volume changes. We calculate this ratio by computing $\frac{D^j_{seg}}{S^j_{seg}}$ where $j = \{0, 1, 2, 3\}$ corresponds to the entire brain and the different tumor classes (NCR/NET, ET, and ED). We then assess the change of the tumor by calculating the absolute value of the difference between $j = 1$ and every other tumor class. Ideally, we expect a model which preserves the tumor geometry and shape during inference to present a zero difference between the entire brain and tumor class ratio. We independently calculate this difference for each tumor class in order to monitor the behavior of each class, but also after merging the entire tumor area.

For statistical significance evaluations between any two methods, we compute independent *t*-tests as presented in Rouder et al. (2009), defining as null hypothesis the evaluation metrics of the two populations to be equal. We then report the associated *p*-value, and the Cohen's d (Rice and Harris, 2005), which we use to measure the effect size. Such statistical significance evaluation is reported in the form ($t(n)$; p; d) where n is the number of samples for each population, $t(n)$ is the t-value, p is the p-value and d is Cohen's d. We defined the difference of two population means is statistically significant if the associated p-value is lower than 0.005, and consider, as a rule of thumb, that a value of d of 0.20 indicates small effect size, 0.50 for medium effect size and 0.80 for large effect size. All of the results in this paper have been computed on unseen testing sets, and the performance of all benchmarked models has been assessed once.

For rigor and for each *t*-test conducted, we ensure the following assumptions are met by the underlying distributions: observations are independent and identically distributed, the outcome variable follows a normal distribution in the population (with Jarque and Bera, 1980), and the outcome variable has equal standard deviations in two considered (sub)populations [using Levene's test (Schultz, 1985)]. Finally, when comparing two populations, each made of several subpopulations, we merge such subpopulations into a single set, then compute *t*-tests on the obtained two gathered-populations.

3. RESULTS

3.1. Evaluation of the Segmentation

Segmentation results for the tumor regions are displayed in **Table 2** for the case of the same autoencoder architecture trained only with a segmentation decoder (Baseline *segmentation*) and the proposed method using different merging operations and with or without $\mathcal{L}^{\star}_{reg}$. One can observe that all evaluated methods perform quite similarly with Dice higher than 0.66 for all the classes and models. The *baseline segmentation* model reports slightly better average Dice coefficient and average Haussdorf distance measurements, with an average Dice 0.03 higher, and an average Hausdorff95 distance 0.6 higher than the proposed with concatenation merging operator, although none of these differences are found statistically significant as indicated in **Table 3**. In particular, for Dice, the minimum received p-value was $p = 0.24$, reported between *baseline segmentation* and *proposed concatenation with* $\mathcal{L}^{\star}_{reg}$ together with an associated Cohen's $d = 0.21$ indicating a small size effect. Similarly, for Hausdorff95, the minimum received p-value was $p = 0.46$, reported this time between *baseline segmentation* and *proposed concatenation w/o* $\mathcal{L}^{\star}_{reg}$ with $d = 0.13$ also indicating a small size effect. These numbers show that the means differences between those two models and any other two models are not statistically significant. This is very promising if we take into account that our proposed model is learning a far more complex architecture addressing both registration and segmentation, with the same volume of training data without significant drop of the segmentation performance.

The superiority of the *baseline segmentation* seems to be presented mainly due to higher performance for the TC class [*baseline segmentation* and *proposed subtraction with* $\mathcal{L}^{\star}_{reg}$: $t_{(66)} = 1.41$; $p = 0.16$; $d = 0.24$]. Moreover, the concatenation operation seems to perform slightly better for the tumor segmentation than the subtraction, with at least 0.02 improvement for average Dice coefficient, although this improvement is not statistically significant [*proposed concatenation with* $\mathcal{L}^{\star}_{reg}$ and *proposed subtraction with* $\mathcal{L}^{\star}_{reg}$: $t_{(66)} = 0.62$; $p = 0.53$; $d = 0.11$].

Moreover, even if one of the main goals of our paper is the proper registration of the tumoral regions, we perform a

TABLE 2 | Quantitative results of the different methods on the segmentation task on the BraTS 2018 validation dataset.

	Average		Dice			Hausdorff95		
Method	**Dice**	**Hausdorff95**	**ET**	**WT**	**TC**	**ET**	**WT**	**TC**
Baseline segmentation	0.79 ± 0.29	7.0 ± 9.6	0.73 ± 0.29	0.87 ± 0.13	0.75 ± 0.24	4.7 ± 8.2	7.2 ± 9.4	9.2 ± 8.9
Proposed								
Concatenation w/o $\mathcal{L}^{\star}_{reg}$	0.74 ± 0.29	8.3 ± 10.4	0.70 ± 0.29	0.87 ± 0.11	0.65 ± 0.29	6.2 ± 9.8	7.8 ± 11.1	11.3 ± 7.1
Concatenation with $\mathcal{L}^{\star}_{reg}$	0.73 ± 0.29	7.6 ± 9.9	0.68 ± 0.30	0.87 ± 0.12	0.66 ± 0.28	6.3 ± 9.9	5.6 ± 4.2	10.8 ± 6.6
Subtraction w/o $\mathcal{L}^{\star}_{reg}$	0.76 ± 0.27	7.8 ± 10.3	0.71 ± 0.28	0.88 ± 0.10	0.70 ± 0.24	6.5 ± 10.8	7.4 ± 11.0	10.0 ± 7.4
Subtraction with $\mathcal{L}^{\star}_{reg}$	0.76 ± 0.27	7.9 ± 10.1	0.71 ± 0.29	0.88 ± 0.10	0.69 ± 0.25	5.8 ± 9.6	7.7 ± 11.5	11.1 ± 8.3

Dice and Hausdorff95 are reported for the three classes Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC) together with their average values. Results are reported with mean across patients (MRIs) along with the associated standard deviation. We upload our predictions on the official leaderboard of the validation set (66 patients).

TABLE 3 | Statistical significance of the proposed methods with Milletari et al. (2016) on the BraTS segmentation task.

Method	Average		Dice			Hausdorff95		
	Dice	Hausdorff95	ET	WT	TC	ET	WT	TC
Baseline segmentation	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Proposed								
Concatenation w/o $\mathcal{L}^{\star}_{reg}$	0.32	0.46	0.55	1.00	0.03	0.34	0.74	0.14
Concatenation with $\mathcal{L}^{\star}_{reg}$	0.24	0.72	0.33	1.00	0.05	0.31	0.21	0.24
Subtraction w/o $\mathcal{L}^{\star}_{reg}$	0.55	0.65	0.69	0.62	0.24	0.28	0.91	0.58
Subtraction with $\mathcal{L}^{\star}_{reg}$	0.55	0.60	0.69	0.62	0.16	0.48	0.79	0.21

For each model (line) and each performance measure (column), the displayed value is the p-value, up to two significant figures, of the statistical significance between the model and Milletari et al. (2016) for the corresponding measure (Dice or Hausdorff95) on the corresponding tumor class (ET, WT, TC, or the union of the three latter in the two columns Average) on the 66 testing samples of BraTS. No p-values are statistically significant between all of the proposed variants and Milletari et al. (2016). Blue line represents the reference model, red cells indicate no statistical significant p-values (cutoff 0.005).

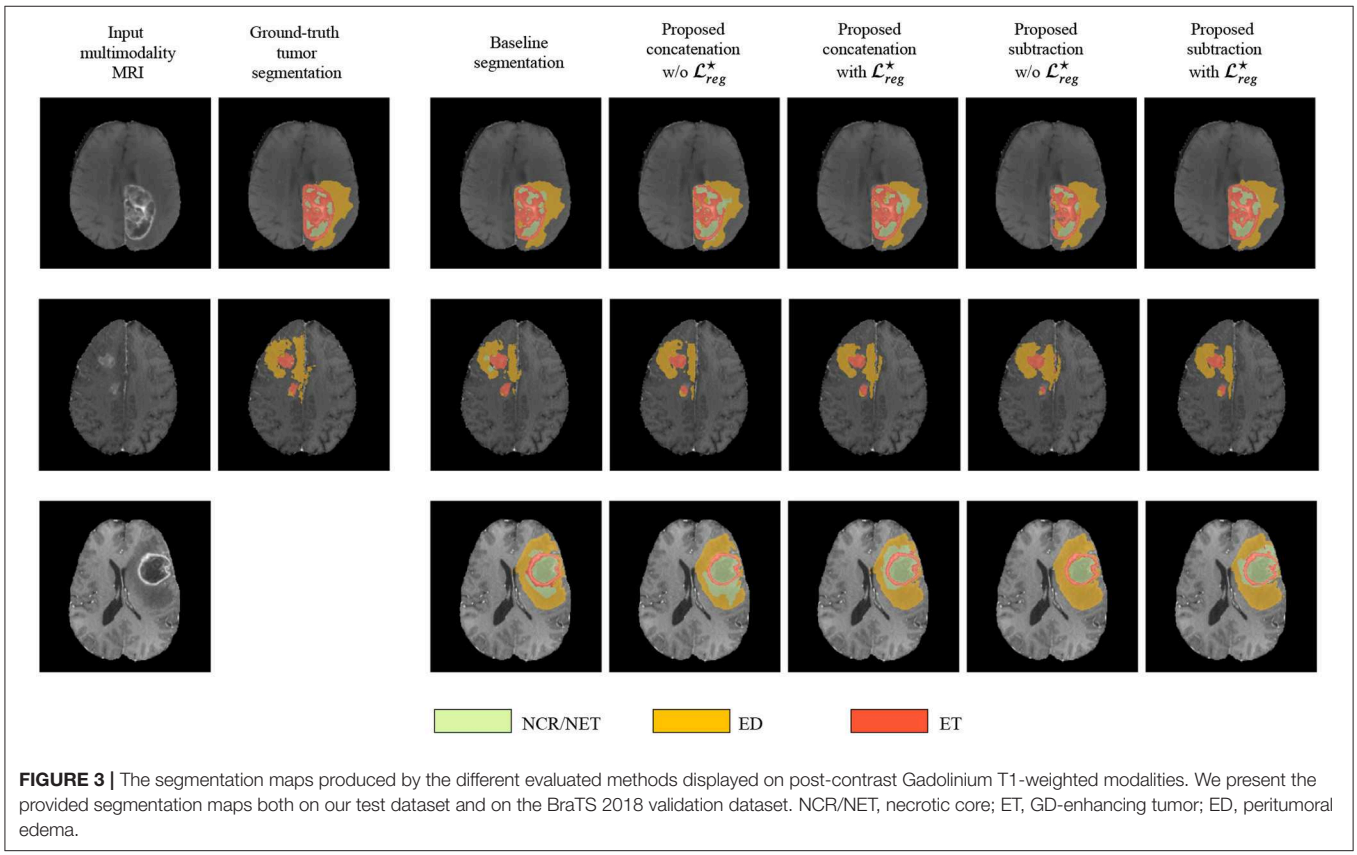


FIGURE 3 | The segmentation maps produced by the different evaluated methods displayed on post-contrast Gadolinium T1-weighted modalities. We present the provided segmentation maps both on our test dataset and on the BraTS 2018 validation dataset. NCR/NET, necrotic core; ET, GD-enhancing tumor; ED, peritumoral edema.

comparison with the two best performing methods presented in BraTS 2018 (Isensee et al., 2018; Myronenko, 2018) evaluated on the validation dataset of BraTS 2018. In particular, the Myronenko (2018) reports an average dice of 0.82, 0.91, and 0.87 for ET, WT, and TC, respectively, while Isensee et al. (2018) reports 0.81, 0.91, and 0.87. Both methods outperform our proposed approach on the validation set of BraTS 2018 by integrating novelties specifically designed to the tumor segmentation task of BraTS 2018. In this study, we based our architecture in a relatively simple and widely used 3D fully convolutional network (Milletari et al., 2016) although different architectures with tumor specific components (trained on the evaluated tumor classes), trained on more data (similar to the ones that are used from Isensee et al., 2018), or even integrating post-processing steps can be easily integrated boosting considerably the performance of our method.

Finally, in **Figure 3** we represent the ground truth and predicted tumor segmentation maps comparing the *baseline segmentation* and our proposed method using the different components and merging operators. We present three different cases, two from our custom test set, on which we have the ground truth information and one from the validation set of the BraTS submission page. One can observe that all the methods

provide quite accurate segmentation maps for all the three tumor classes.

3.2. Evaluation of the Registration

3.2.1. Evaluation on Anatomical Structures

The performance of the registration has been evaluated on an unseen dataset with anatomical information, namely OASIS 3. In **Table 4** the mean and standard deviation of the Dice coefficient for the different evaluated methods are presented. With rigid we indicate the Dice coefficient after the translation of the volumes such that the center of the brain mass is placed in the center of the volume. It can be observed that the performance of the evaluated methods are quite similar something which indicates that the additional tumor segmentation decoder does not decrease the performance of the registration. On the other hand, it provides additional information about the areas of tumor in the image. From our experiments, we show that the proposed formulation can provide registration accuracy similar to the recent state-of-the-art deep learning based methods (Dalca et al., 2018) with approximately the same average Dice values, that is 0.50 for (Dalca et al., 2018) and 0.49 for all but one of the proposed variants. Moreover, again this difference in the performance between (Dalca et al., 2018) and the proposed method is not statistically significant with $t_{(150)} = 0.64$; $p = 0.52$; $d = 0.07$. From our comparisons, the only significant difference on the evaluation of the registration task was reported between the proposed method *concatenation only reg.* with an average difference of dice reaching 0.05% and with maximum p-values calculated with *Proposed concatenation with* $\mathcal{L}^{\star}_{reg}$ [$t_{(200)} = 3,33$; $p < 10^{-3}$; $d = 0,38$]. From our experiments, we saw that the merging operation affects the performance of the *only reg.* model a lot, with the concatenation reporting the worst average dice of all the methods.

In **Figure 4** we present some qualitative evaluation of the registration component, by plotting three different pairs and their registration from all the evaluated models. The first two columns of the figure depict the source and reference volumes together with their tissue annotations. The rest of the columns present the deformed source volume together with the deformed tissue annotations for each of the evaluates methods. Visually, all methods perform well on the overall shape of the brain with the higher errors in the deformed annotations being presented at the cerebral write matter and cerebral cortex classes.

Finally, we should also mention that the subjects of the OASIS 3 dataset do not contain regions with tumors. However, our proposed formulation provides tumor masks so that we could evaluate the robustness of the segmentation part. Indeed, our model for all the different combinations of merging operations and loss functions, reported a precision score of more than 0.999, indicating its robustness for the tumor segmentation task.

3.2.2. Evaluation on the Tumor Areas

Even if the proposed method reports very similar performance with models that perform only registration, we argue that it addresses better the registration of the tumor areas, maintaining their geometric properties, as can be inferred in **Table 5**. This statement is also supported by the statistical tests we performed

TABLE 4 | The mean and standard deviation of the dice coefficient for the 15 different classes of OASIS 3 dataset for the different evaluated methods.

Method	BS	CSF	CblmC	CblmWM	CeblWM	Pu	VDC	Pa	Ca	LV	Hi	3V	4V	Am	CebIC	Average
Rigid	0.58 ± 0.15	0.39 ± 0.11	0.46 ± 0.13	0.40 ± 0.14	0.49 ± 0.05	0.44 ± 0.13	0.47 ± 0.13	0.35 ± 0.17	0.27 ± 0.15	0.40 ± 0.13	0.34 ± 0.13	0.39 ± 0.17	0.15 ± 0.15	0.24 ± 0.18	0.36 ± 0.04	0.38 ± 0.13
Voxelmorph	0.69 ± 0.12	**0.46 ± 0.13**	**0.63 ± 0.11**	**0.57 ± 0.13**	**0.73 ± 0.083**	0.42 ± 0.14	0.5 ± 0.11	0.33 ± 0.14	0.42 ± 0.17	0.62 ± 0.14	0.38 ± 0.13	**0.53 ± 0.18**	**0.32 ± 0.23**	0.25 ± 0.17	**0.6 ± 0.084**	**0.5 ± 0.14**
Proposed																
Concatenation																
Only reg.	0.65 ± 0.15	0.34 ± 0.1	0.58 ± 0.11	0.48 ± 0.14	0.6 ± 0.056	0.46 ± 0.12	0.47 ± 0.12	0.38 ± 0.14	0.35 ± 0.15	0.54 ± 0.14	0.35 ± 0.13	0.4 ± 0.16	0.21 ± 0.17	0.27 ± 0.18	0.46 ± 0.051	0.44 ± 0.13
w/o $\mathcal{L}^{\star}_{reg}$	0.72 ± 0.13	0.42 ± 0.1	0.61 ± 0.11	0.51 ± 0.12	0.63 ± 0.056	0.47 ± 0.14	0.51 ± 0.12	0.37 ± 0.16	**0.44 ± 0.15**	**0.65 ± 0.13**	**0.42 ± 0.14**	0.46 ± 0.17	0.31 ± 0.22	0.31 ± 0.19	0.48 ± 0.052	0.49 ± 0.13
With $\mathcal{L}^{\star}_{reg}$	0.7 ± 0.15	0.44 ± 0.12	0.6 ± 0.13	0.52 ± 0.14	0.66 ± 0.06	0.47 ± 0.14	0.52 ± 0.13	0.38 ± 0.16	0.42 ± 0.16	0.65 ± 0.14	0.4 ± 0.15	0.51 ± 0.19	0.3 ± 0.22	0.28 ± 0.2	0.49 ± 0.058	0.49 ± 0.14
Subtraction																
Only reg.	0.71 ± 0.13	0.41 ± 0.1	0.61 ± 0.12	0.53 ± 0.13	0.66 ± 0.058	0.47 ± 0.12	0.5 ± 0.11	0.37 ± 0.15	0.43 ± 0.14	0.63 ± 0.12	0.4 ± 0.13	0.47 ± 0.16	0.34 ± 0.22	0.29 ± 0.19	0.49 ± 0.054	0.49 ± 0.13
w/o $\mathcal{L}^{\star}_{reg}$	0.7 ± 0.13	0.41 ± 0.1	0.6 ± 0.11	0.52 ± 0.12	0.65 ± 0.057	**0.48 ± 0.13**	**0.53 ± 0.11**	**0.39 ± 0.15**	0.43 ± 0.14	0.64 ± 0.13	0.41 ± 0.13	0.49 ± 0.17	0.3 ± 0.22	0.29 ± 0.18	0.48 ± 0.053	0.49 ± 0.13
With $\mathcal{L}^{\star}_{reg}$	**0.72 ± 0.12**	0.4 ± 0.11	0.61 ± 0.11	0.53 ± 0.12	0.64 ± 0.058	0.47 ± 0.12	0.51 ± 0.11	0.38 ± 0.15	0.41 ± 0.15	0.63 ± 0.13	0.43 ± 0.13	0.44 ± 0.17	0.3 ± 0.22	**0.33 ± 0.18**	0.48 ± 0.054	0.49 ± 0.13

The first two rows are baseline methods. The rest of the rows present the results of our proposed method evaluating the different variants and merging operators. The names of the columns represent various brain structures, namely: brain stem (BS), cerebrospinal fluid (CSF), 4th ventricle (4V), amygdala (Am), caudate (Ca), cerebellum cortex (CblmC), cerebellum white matter (CblmWM), cerebral cortex (CebIC), cerebral white matter (CeblWM), hippocampus (Hi), lateral ventricle (LV), pallidum (Pa), putamen (Pu), ventral DC (VDC), and 3rd ventricle (3V). Bold indicates best performance per column.

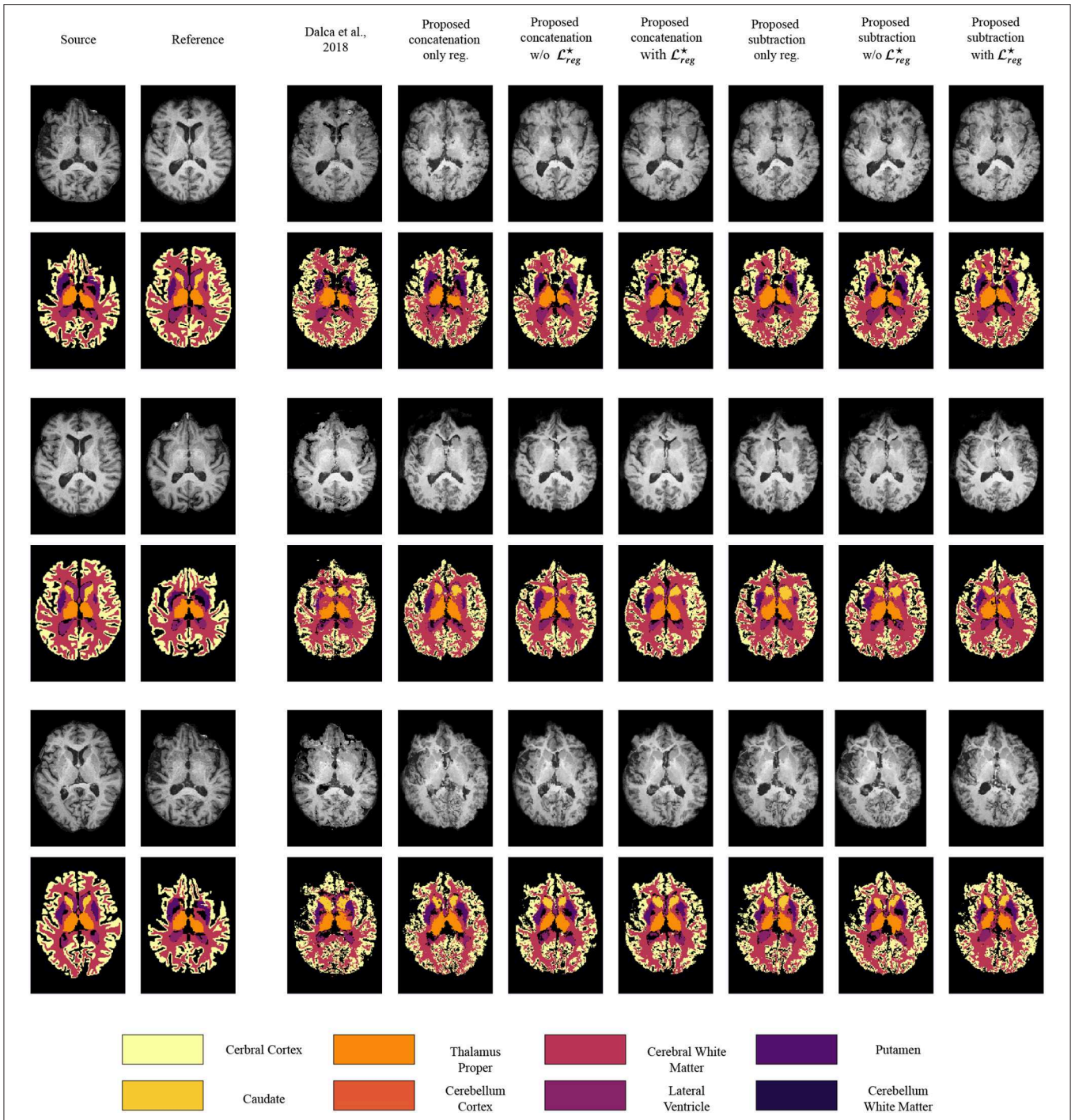


FIGURE 4 | Qualitative evaluation of the registration performance for the different evaluated methods, displayed on T1 modalities. For an easier visualization, we group left and right categories and only display the following nine classes: caudate (Ca), cerebellum cortex (CblmC), cerebellum white matter (CblmWM), cerebral cortex (CeblC), cerebral white matter (CeblWM), lateral ventricle (LV), pallidum (Pa), putamen (Pu), ventral DC (VDC).

to evaluate the difference in performance between the methods, while registering tumor areas (**Table 6**). In particular, for each of the tumor classes NCR/NET, ET, and ED the difference between (Dalca et al., 2018) and the proposed method *subtraction with* $\mathcal{L}^{\star}_{reg}$ was significant with NCR/NET: $t_{(200)} = 10.69$; $p < 10^{-3}$; $d = 1.07$—ET: $t_{(200)} = 10.51$; $p < 10^{-3}$; $d = 1.05$—ED: $t_{(200)} = 8.05$; $p < 10^{-3}$; $d = 0.81$. The similar behavior was obtained when the evaluation was performed by merging

TABLE 5 | The table presents the average distance between (i) the ratio of the area of the deformed tumor mask to the area of the original tumor mask, and (ii) the ratio of area of the reference brain volume to the area of the source brain volume.

Method	NCR/NET	ET	ED	Combined
Dalca et al. (2018)	2.27 ± 2.68	0.67 ± 0.55	1.96 ± 3.03	0.62 ± 0.51
Proposed				
Concatenation only reg.	0.51 ± 0.61	0.26 ± 0.19	0.71 ± 0.94	0.22 ± 0.15
Concatenation w/o $\mathcal{L}^{\star}_{reg}$	1.35 ± 1.14	0.64 ± 0.41	1.80 ± 1.82	0.64 ± 0.42
Concatenation with $\mathcal{L}^{\star}_{reg}$	0.26 ± 0.20	0.26 ± 0.13	0.30 ± 0.28	0.21 ± 0.12
Subtraction only reg.	1.34 ± 0.77	0.77 ± 0.59	2.02 ± 1.65	0.68 ± 0.52
Subtraction w/o $\mathcal{L}^{\star}_{reg}$	1.74 ± 1.35	0.72 ± 0.72	2.38 ± 1.74	0.74 ± 0.76
Subtraction with $\mathcal{L}^{\star}_{reg}$	**0.24 ± 0.17**	**0.25 ± 0.13**	**0.23 ± 0.22**	**0.20 ± 0.11**

Lower values are better. The average has been calculated over 200 testing pairs from the BraTS 2018 dataset (NCR/NET, ET and ED). On top of the evaluation per tumor class, we also conduct an evaluation by merging all the tumor classes into just one class (called combined). Bold indicates best performance per column.

TABLE 6 | Summary of the statistical difference between the Dalca et al. (2018) and the proposed method on the BraTS 2018 dataset for the tumor preservation task.

Method	NCR/NET	ET	ED	Combined
Dalca et al. (2018)	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Proposed				
Concatenation only reg.	$< 10^{-3}$	0.540	$< 10^{-3}$	0.130
Concatenation w/o $\mathcal{L}^{\star}_{reg}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Concatenation with $\mathcal{L}^{\star}_{reg}$	0.282	0.442	0.006	0.386
Subtraction only reg.	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Subtraction w/o $\mathcal{L}^{\star}_{reg}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Subtraction with $\mathcal{L}^{\star}_{reg}$	1.000	1.000	1.000	1.000

For each model (line) and each performance measure (column), the displayed value is the p-value (up to 3 significant figures) of the statistical significance between the model and subtraction with $\mathcal{L}^{\star}_{reg}$ for the tumor preservation measure on the corresponding tumor class (NCR/NET, ET, ED, and their union in the column Combined). Blue line represents the reference model, red cells indicate no statistical significant p-values while green cells represents statistical significant p-values.

all 3 tumor classes into one (denoted *Combined*). Again, we reported significant differences between (Dalca et al., 2018) and the proposed method: $t_{(200)} = 11.38; p < 10^{-3}; d = 1.14$.

To evaluate the performance of the different variants of our proposed method, we compared the performance of the proposed *subtraction with* $\mathcal{L}^{\star}_{reg}$ and *concatenation with* $\mathcal{L}^{\star}_{reg}$ that reported the best performances. Indeed, we did not find significant changes between the two different components except the edema class [$t_{(200)} = 2.78; p < 10^{-3}; d = 0.28$]. Moreover, the proposed *concatenation only reg.* reports also competitive results without using the segmentation masks. In particular, even if the specific method does not report very good performance on the registration evaluated on anatomical structures (section 3.2.1), it reports very competitive performance on the *Combined* and the smallest in size tumor class (*ET*). However, for the other two classes the difference on the performance that it reports in comparison to the proposed variant *subtraction with* $\mathcal{L}^{\star}_{reg}$ is significant different: NCR/NET: $t_{(200)} = 6,03; p < 10^{-3}; d = 0,60$—ED: $t_{(200)} = 7,03; p < 10^{-3}; d = 0,70$. Here we should mention that even though *subtraction only reg.* works very well for the registration of the anatomical regions (section 3.2.1), it reports one of the worst results about tumor preservation, with values close to the ones reported by Dalca et al. (2018). This indicates again that the *only reg.* model is highly sensitive to the merging operation and it cannot simultaneously provide good performance on tumor areas and registration of the entire volume, proving its inferiority to the proposed method using the *with* $\mathcal{L}^{\star}_{reg}$.

Independently of the merging operation with both registration and segmentation tasks, i.e., with or without $\mathcal{L}^{\star}_{reg}$, we find that the proposed approach works significantly better in preserving tumor areas when optimized with $\mathcal{L}^{\star}_{reg}$ than without [NCR/NET: $t_{(200)} = -14.33; p < 0.005; d = 1.43$—ET: $t_{(200)} = -9.99; p < 0.005; d = 1.00$—ED: $t_{(200)} = -14.17; p < 0.005; d = 1.42$—Combined: $t_{(200)} = -10.94; p < 0.005; d = 1.09$].

Figure 5 presents some qualitative examples from the BraTS 2018 to evaluate the performance of the different methods. The first two columns present the pair of images to be registered and segmented and the rest of the columns the deformed source image with the segmented tumor region superimposed. One can observe that the most of the methods that are based only on registration (Dalca et al., 2018, proposed concatenation and subtraction *only reg.*) together with the proposed concatenation and subtraction *w/o* $\mathcal{L}^{\star}_{reg}$ do not preserve the geometry of the tumor, tending to significantly reduce the area of tumor after registration, or intermix the different types of tumor. On the other hand the behavior of the proposed *with* $\mathcal{L}^{\star}_{reg}$ seems to be much better, with the tumor area properly maintained in the deformed volume.

Moreover, in **Figure 6** we provide a better visualization for the displacement grid inside the tumor area, highlighting the importance of Equation (3). Indeed, one can observe that the displacements inside the tumor area are much smoother and relaxed when we use the information about the tumor segmentation.

4. DISCUSSION

In this study, we proposed a novel deep learning based framework to address simultaneously segmentation and registration. The framework combines and generates features, integrating valuable information from both tasks within a bidirectional manner, while it takes advantage of all the available modalities, making it quite robust and generic. The performance of our model indicates highly promising results that are comparable to recent state-of-the-art models that address each of the tasks separately (Dalca et al., 2018). However, we reported a better behavior of the model in the proximity of tumor regions. This behavior has been achieved by training a shared encoder that generates features that are meaningful for both registration and segmentation problems. At the same time, these two problems have been coupled in a joint loss

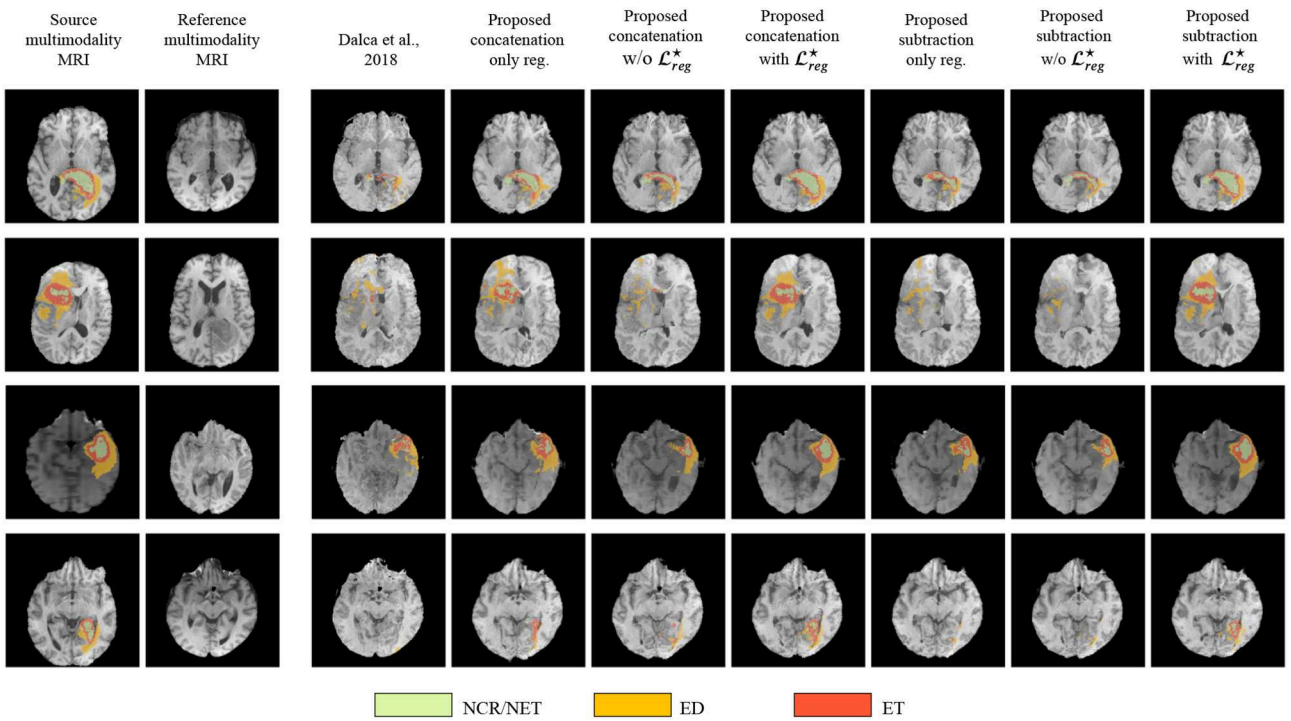


FIGURE 5 | Qualitative evaluation of the tumor deformation of the different evaluated methods, displayed on T1 modalities. Each line is a sample, with source MRI in the first column to be registered on reference MRI in the second column. BraTS ground-truth annotations are plotted onto the source MRI. Seven models are benchmarked, one for each of the remaining columns which display the result of applying the predicted grid onto the source MRI. For each model and each line, the source ground-truth annotation masks of the source MRI were also registered with the predicted deformation grid, and the consequently obtained deformed ground-truth were plotted onto each deformed source MRI to illustrate the impact of all methods regarding the preservation of tumor extent.

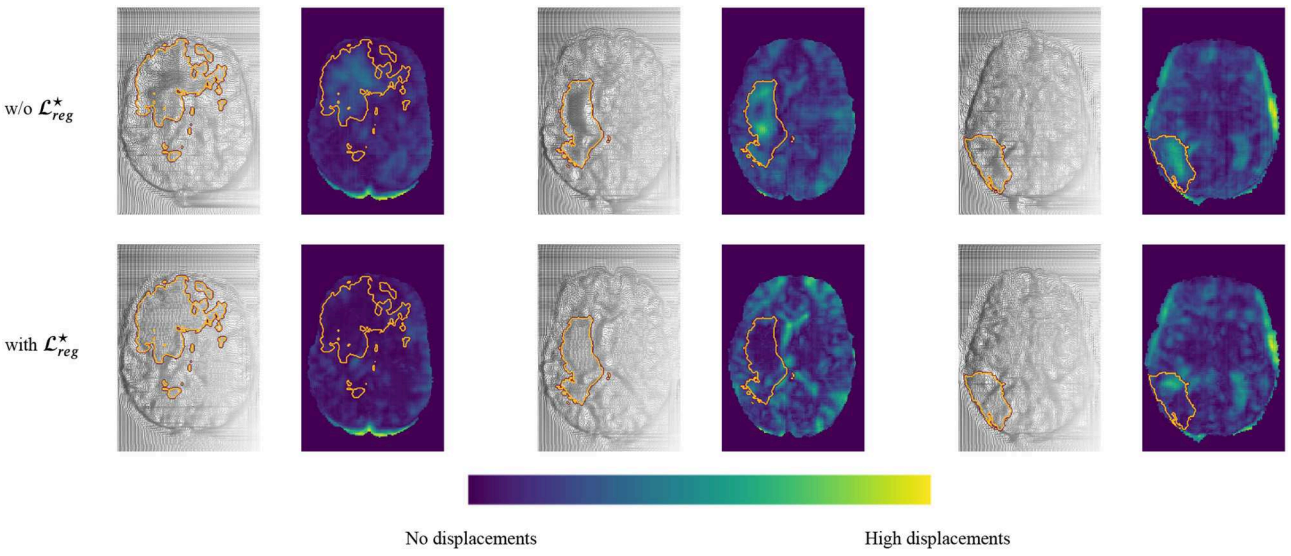


FIGURE 6 | Comparison of the registration grid of the proposed model using the subtraction operation with and w/o $\mathcal{L}^{\star}_{reg}$. This figure is obtained by sampling three random pairs of test patients, and computing the predicted registration fields, which are displayed by line for the two models, and in consecutive columns, one for each of the three dimensions, showing the registration field as a warped grid (grayscale) and as a colored map obtained by computing its norm pixelwise (blue-green map). Furthermore, the contour of the Whole Tumor is plotted on top of each image, obtained from the ground truth segmentation.

function, enforcing the network to focus on regions that exist in both volumes.

Even if we could not do a proper comparison with Parisot et al. (2012) which shares similar concepts, our method provides very good improvements. In particular, we train both problems at the same time, without using pre-calculated classification probabilities. The method proposed in Parisot et al. (2012) is based on a pre-calculated classifier indicating the tumoral regions. The authors provided their segmentation results by adapting Gentle Adaboost algorithm and using different features including intensity values, texture, such as Gabor filters and symmetry. After training the classifier they defined an MRF model to optimize their predictions by taking into account pairwise relations. By adopting this strategy, the used probabilities for the tumoral regions are not optimized simultaneously with the registration, something that it is not the case in our methodology. In particular, by sharing representation between the registration and segmentation tasks we argue that we can create features that are more complex and useful sharing information that comes from both problems. By using a deep learning architecture that is end-to-end trainable, we are able to extract features that are suitable to deal with both problems automatically. Moreover, our implementation is modular and scalable permitting easy integration of multiple modalities, something that is not so straightforward with Parisot et al. (2012) as it is more complicated to adapt and calculate the different similarity measures and classifiers taking into account all these modalities. Finally, we should mention that our method takes advantage of GPU implementation needing only a few seconds in order to provide segmentation and displacement maps while the method in Parisot et al. (2012) needs ∼6 min.

Both qualitative and quantitative evaluations of the proposed architecture highlight its great potentials reporting more than 0.66 Dice coefficient for the segmentation of the different tumor areas, evaluated on the publicly available BraTS 2018 validation set. Our joint formulation reported performance similar to the model trained only for segmentation, while simultaneously addressing the registration problem. Moreover, both concatenation and subtraction operators report similar performances, an expected result for the specific segmentation task, since the merging operation is mainly used on the registration decoder, even if it affects the learned parameters of the encoder and thus indirectly the segmentation decoder.

Concerning the comparison between top performing tumor segmentation methods, although our formulation underperforms the winning methods of BraTS 2018, we want to highlight two major points. First of all, our formulation is modular in the sense that different network architectures with optimized components for tumor segmentation can be evaluated depending on the application and the goals of the problem. For our experiments we chose a simple VNet architecture (Milletari et al., 2016) proving that the registration components do not significantly hinder the segmentation performance and indicating the soundness of our method however any other encoder decoder architecture can be used and evaluated. Secondly, the main goal of our method was the proper registration and segmentation of the tumoral regions together with the rest of the anatomical structures and that was the main reason we did not optimize our network architecture according to the winning methods of BraTS 2018. However, we demonstrated that with a very simple architecture, we can register properly tumoral and anatomical structures while segmenting with more than 76% of Dice the tumoral regions.

Continuing with the evaluation of the registration performance, once more the joint multi task framework reports similar and without statistical difference performance with formulations that address only the registration task evaluated on anatomical regions that exist on both volumes. However, we argue that abnormal regions registration is better addressed both in terms of qualitative and quantitative metrics. Moreover, from our experiments we observed that subtraction of the coding features of the tumors reports higher performances for the registration of the tumor areas. This indicates that the subtraction can capture and code more informative features for the registration task. What is more, we achieved very good generalization for all the deep learning based registration methods, as they reported very stable performance in a completely unseen dataset (part of the OASIS3).

Even if, from our experiments, the competence of our proposed method for both registration and segmentation tasks is indicated, we report a much better performance for the registration of the tumoral regions. In particular, in one joint framework we were able to produce efficiently and accurately tumor segmentation maps for both source and reference images together with their displacement maps that register the source volume to the reference volume space. Our experiments indicated that the proposed method with the $\mathcal{L}^{\star}_{reg}$ variant register properly the anatomical together with the tumoral regions with statistical significance compare to the rest of the methods for the latter. Both qualitative and quantitative evaluations of the different components indicate the superiority of the with $\mathcal{L}^{\star}_{reg}$ variant of the proposed method for brain MRI registration with tumor extent preservation. Using such a formulation, the network focus on improving local displacements on tissues anywhere in the common brain space instead of minimizing the loss within the tumoral regions, which are empirically the regions with the highest registration errors. Consequently, the network improves its registration performance on non-tumor regions (as discussed in section 3.2.1), while also relaxing the obtained displacements inside those predicted tumor regions.

Some limitations of our method include the number of parameters that have to be tuned during the training due to the multi task nature of our formulation, namely α and β that affect the performance of the network. Moreover, due to the multimodal nature of the input and the two decoders, the network cannot be very deep due to GPU memory limitations.

Although the pipeline was built using different patients for the registration task as a proof of concept, such tool could have numerous applications in clinical practice, especially when applied in different images acquired from the same patient. Regarding the radiotherapy treatment planning, several studies have shown that significant changes of the targeted volumes in the brain occurred during radiotherapy raising the question of replanning treatment to reduce the amount of healthy brain irradiated in case of tumor reduction, or to re-adapt the treatment for brain tumors that grow during radiation (Champ et al., 2012; Yang et al., 2016; Mehta et al., 2018). Since MR-guided linear accelerator will offer the opportunity to acquire daily images during RT treatment, the proposed tool could help with automatic segmentation and image registration for replanning purposes, and it could also allow accurate evaluation of the dose delivered in targeted volumes and healthy tissues by taking into account the different volume changes. Moreover, while changes of imaging features under treatment is known to be associated with treatment outcomes in several cancer diseases (Vera et al., 2014; Fave et al., 2017), the registration grid computed from two same-patient acquisitions realized at different times allows an objective and precise evaluation of the tumor changes.

Future work involves a better modeling of the prior knowledge through a more appropriate geometric modeling of tumor proximity that encodes more accurately the registration errors in these areas. This modeling can be integrated into the existing formulation with some additions specific to tumor losses that will further constrain its change. Moreover, we have noticed that the use of Fobenius norm during the training of the registration part is very sensitive to artifacts in the volume, preventing the network process from being completely robust. In the future, we aim to evaluate the performance of the proposed framework using adversarial losses in order to better address multimodal cases. Finally, means to automatically obtain the training parameters α and β would be investigated.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2018/data.html, https://www.oasis-brains.org/.

AUTHOR CONTRIBUTIONS

TE, ML, MV, NP, and ED designed the research. TE, ML, and MV performed the research, analyzed and interpreted the data, and wrote the paper. TE, ML, MV, EA, EB, AC, SCha, SChr, MS, RS, CR, HT, NP, and ED revised and approved the paper.

FUNDING

This work have been partially funding by the ARC: Grant SIGNIT201801286, the Fondation pour la Recherche Médicale: Grant DIC20161236437, SIRIC-SOCRATE 2.0, ITMO Cancer, Institut National du Cancer (INCa), and Amazon Web Services (AWS).

ACKNOWLEDGMENTS

We would like to acknowledge Y. Boursin, M. Azoulay, and Gustave Roussy Cancer Campus DTNSI team for providing the infrastructure resources used in this work as well as Amazon Web Services for their partial support.

REFERENCES

Avants, B., Epstein, C., Grossman, M., and Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Archiv.* doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017c). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. *Cancer Imaging Archiv.* doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv[Prepeint]*.arXiv:1811.02629

Champ, C. E., Siglin, J., Mishra, M. V., Shen, X., Werner-Wasik, M., Andrews, D. W., et al. (2012). Evaluating changes in radiation treatment volumes from post-operative to same-day planning mri in high-grade gliomas. *Radiat. Oncol.* 7:220. doi: 10.1186/1748-717X-7-220

Chandra, S., Vakalopoulou, M., Fidon, L., Battistella, E., Estienne, T., Sun, R., et al. (2019). "Context aware 3D CNNs for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 299–310.

Christodoulidis, S., Sahasrabudhe, M., Vakalopoulou, M., Chassagnon, G., Revel, M.-P., Mougiakakou, S., et al. (2018). "Linear and deformable image registration with 3D convolutional neural networks," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, eds D. Stoyanov, Z. Taylor, B. Kainz, G. Maicas, R. R. Beichel, A. Martel, L. Maier-Hein, K. Bhatia, T. Vercauteren, O. Oktay, G. Carneiro, A. P. Bradley, J. Nascimento, H. Min, M. S. Brown, C. Jacobs, B. Lassen-Schmidt, K. Mori, J. Petersen, R. San José Estépar, A. Schmidt-Richberg, C. Veiga (Cham: Springer International Publishing), 13–22.

Dalca, A. V., Balakrishnan, G., Guttag, J. V., and Sabuncu, M. R. (2018). "Unsupervised learning for fast probabilistic diffeomorphic registration," in *MICCAI* (Cham).

Fave, X., Zhang, L., Yang, J., Mackin, D., Balter, P., Gomez, D., et al. (2017). Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer. *Sci. Rep.* 7:588. doi: 10.1038/s41598-017-00665-z

Glocker, B., Komodakis, N., Navab, N., Tziritas, G., and Paragios, N. (2009). "Dense registration with deformation priors," in *Information Processing in Medical Imaging*, eds J. L. Prince, D. L. Pham, and K. J. Myers (Berlin; Heidelberg: Springer), 540–551.

Gooya, A., Biros, G., and Davatzikos, C. (2010). Deformable registration of glioma images using em algorithm and diffusion reaction modeling. *IEEE Trans. Med. Imaging* 30, 375–390. doi: 10.1109/TMI.2010.2078833

Gooya, A., Pohl, K. M., Bilello, M., Cirillo, L., Biros, G., Melhem, E. R., et al. (2012). Glistr: glioma image segmentation and registration. *IEEE Trans. Med. Imaging* 31, 1941–1954. doi: 10.1109/TMI.2012.2210558

Holland, E. C. (2002). Progenitor cells and glioma formation. *Curr Opin Neurol.* 14, 683–688. doi: 10.1097/00019052-200112000-00002

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2017). "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 287–297.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Springer), 234–244.

Jarque, C. M., and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity, and serial independence of regression residuals. *Econ. Lett.* 6, 255–259.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2018). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 450–462.

Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. W. (2009). Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* 29, 196–205. doi: 10.1109/TMI.2009.2035616

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* 22, 2677–2684. doi: 10.1162/jocn.2009.21407

McKinley, R., Meier, R., and Wiest, R. (2018). "Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 456–465.

Mehta, S., Gajjar, S. R., Padgett, K. R., Asher, D., Stoyanova, R., Ford, J. C., et al. (2018). Daily tracking of glioblastoma resection cavity, cerebral edema, and tumor volume with MRI-guided radiation therapy. *Cureus* 10:e2346. doi: 10.7759/cureus.2346

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34. doi: 10.1109/TMI.2014.2377694

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA: IEEE), 565–571.

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 311–320.

Ou, Y., Sotiras, A., Paragios, N., and Davatzikos, C. (2011). Dramms: deformable registration via attribute matching and mutual-saliency weighting. *Med. Image Anal.* 15, 622–639. doi: 10.1016/j.media.2010.07.002

Parisot, S., Darlix, A., Baumann, C., Zouaoui, S., Yordanova, Y., Blonski, M., et al. (2016). A probabilistic atlas of diffuse who grade II glioma locations in the brain. *PLoS ONE* 11:e0144200. doi: 10.1371/journal.pone.0144200

Parisot, S., Duffau, H., Chemouny, S., and Paragios, N. (2012). "Joint tumor segmentation and dense deformable registration of brain MR images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012*, eds N. Ayache, H. Delingette, P. Golland, and K. Mori (Berlin; Heidelberg: Springer), 651–658.

Postelnicu, G., Zollei, L., and Fischl, B. (2009). Combined volumetric and surface registration. *IEEE Trans. Med. Imaging* 28, 508–522. doi: 10.1007/978-3-642-33418-4_80

Rice, M. E., and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and R. *Law Hum. Behav.* 29, 615–620. doi: 10.1007/s10979-005-6832-7

Robinson, E. C., Garcia, K., Glasser, M. F., Chen, Z., Coalson, T. S., Makropoulos, A., et al. (2018). Multimodal surface matching with higher-order smoothness constraints. *Neuroimage* 167, 453–465. doi: 10.1016/j.neuroimage.2017.10.037

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/PBR.16.2.225

Schultz, B. B. (1985). Levene's test for relative variation. *Syst. Zool.* 34, 449–456.

Sepúlveda-Sánchez, J., Langa, J. M., Arráez, M., Fuster, J., Laín, A. H., Reynés, G., et al. (2018). Seom clinical guideline of diagnosis and management of low-grade glioma (2017). *Clin. Transl. Oncol.* 20, 3–15. doi: 10.1007/s12094-017-1790-3

Shi, W., Jantsch, M., Aljabar, P., Pizarro, L., Bai, W., Wang, H., et al. (2013). Temporal sparse free-form deformations. *Med. Image Anal.* 17, 779–789. doi: 10.1016/j.media.2013.04.010

Shu, Z., Sahasrabudhe, M., Riza, A. G., Samaras, D., Paragios, N., and Kokkinos, I. (2018). "Deforming autoencoders: unsupervised disentangling of shape and appearance," in *The European Conference on Computer Vision (ECCV)* (Cham).

Sotiras, A., Davatzikos, C., and Paragios, N. (2013). Deformable medical image registration: a survey. *IEEE Trans. Med. Imaging* 32, 1153–1190. doi: 10.1109/TMI.2013.2265603

Stupp, R., Brada, M., Van Den Bent, M., Tonn, J.-C., and Pentheroudakis, G. (2014). High-grade glioma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 25, iii93–iii101. doi: 10.1093/annonc/mdu050

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, eds M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. Tavares, M. Moradi, et al. (Cham: Springer International Publishing), 240–248.

Vera, P., Dubray, B., Palie, O., Buvat, I., Hapdey, S., Modzelewski, R., et al. (2014). Monitoring tumour response during chemo-radiotherapy: a parametric method using FDG-PET/CT images in patients with oesophageal cancer. *EJNMMI Res.* 4:12. doi: 10.1186/2191-219X-4-12

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 178–190.

Wee, C. W., Sung, W., Kang, H.-C., Cho, K. H., Han, T. J., Jeong, B.-K., et al. (2015). Evaluation of variability in target volume delineation for newly diagnosed glioblastoma: a multi-institutional study from the Korean Radiation Oncology Group. *Radiat. Oncol.* 10:137. doi: 10.1186/s13014-015-0439-z

Yang, Z., Zhang, Z., Wang, X., Hu, Y., Lyu, Z., Huo, L., et al. (2016). Intensity-modulated radiotherapy for gliomas: dosimetric effects of changes in gross tumor volume on organs at risk and healthy brain tissue. *Oncotargets Ther.* 9:3545. doi: 10.2147/OTT.S100455

Zhao, Z., Yang, G., Lin, Y., Pang, H., and Wang, M. (2018). Automated glioma detection and segmentation using graphical models. *PLoS ONE* 13:e0200745. doi: 10.1371/journal.pone.0200745

Zhou, C., Chen, S., Ding, C., and Tao, D. (2018). "Learning contextual and attentive information for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Springer), 497–507.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 08 April 2020
doi: 10.3389/fncom.2020.00025

Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features

Xue Feng[1]*, *Nicholas J. Tustison*[2], *Sohil H. Patel*[2] and *Craig H. Meyer*[1,2]

[1] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, United States, [2] Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA, United States

Accurate segmentation of different sub-regions of gliomas such as peritumoral edema, necrotic core, enhancing, and non-enhancing tumor core from multimodal MRI scans has important clinical relevance in diagnosis, prognosis and treatment of brain tumors. However, due to the highly heterogeneous appearance and shape of these tumors, segmentation of the sub-regions is challenging. Recent developments using deep learning models has proved its effectiveness in various semantic and medical image segmentation tasks, many of which are based on the U-Net network structure with symmetric encoding and decoding paths for end-to-end segmentation due to its high efficiency and good performance. In brain tumor segmentation, the 3D nature of multimodal MRI poses challenges such as memory and computation limitations and class imbalance when directly adopting the U-Net structure. In this study we aim to develop a deep learning model using a 3D U-Net with adaptations in the training and testing strategies, network structures, and model parameters for brain tumor segmentation. Furthermore, instead of picking one best model, an ensemble of multiple models trained with different hyper-parameters are used to reduce random errors from each model and yield improved performance. Preliminary results demonstrate the effectiveness of this method and achieved the 9th place in the very competitive 2018 Multimodal Brain Tumor Segmentation (BraTS) challenge. In addition, to emphasize the clinical value of the developed segmentation method, a linear model based on the radiomics features extracted from segmentation and other clinical features are developed to predict patient overall survival. Evaluation of these innovations shows high prediction accuracy in both low-grade glioma and glioblastoma patients, which achieved the 1st place in the 2018 BraTS challenge.

Keywords: brain tumor segmentation, ensemble, 3D U-net, deep learning, survival prediction, linear regression

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Xin Tian,
Tianjin Medical University, China
Prateek Prasanna,
Stony Brook University, United States

***Correspondence:**
Xue Feng
xf4j@virginia.edu

Received: *31 July 2019*
Accepted: *17 March 2020*
Published: *08 April 2020*

Citation:
Feng X, Tustison NJ, Patel SH and Meyer CH (2020) Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. Front. Comput. Neurosci. 14:25. doi: 10.3389/fncom.2020.00025

INTRODUCTION

Gliomas are the most common primary brain malignancies, with different degrees of aggressiveness, variable prognosis and various heterogeneous histological sub-regions, i.e., peritumoral edema, necrotic core, enhancing, and non-enhancing tumor core (Wrensch et al., 2002; Louis et al., 2016). This intrinsic heterogeneity of gliomas is also portrayed in their radiographic

phenotypes, as their sub-regions are depicted by different intensity profiles disseminated across multimodal MRI (mMRI) scans, reflecting differences in tumor biology (Cha, 2006; Upadhyay and Waldman, 2011). Quantitative analysis of imaging features such as volumetric measures after manual/semi-automatic segmentation of the tumor region has shown advantages in image-based tumor phenotyping over traditionally used clinical measures such as largest anterior-posterior, transverse, and inferior-superior tumor dimensions on a subjectively-chosen slice (Kumar et al., 2012; Gillies et al., 2016). Such phenotyping may enable assessment of reflected biological processes and assist in surgical and treatment planning. For brain tumors, including sub-regions, segmentation is challenging due to their highly heterogeneous appearance and shape, which may be further complicated by imaging artifacts such as motion and/or field inhomogeneity.

In recent years, deep convolutional neural networks (DCNN) have demonstrated effectiveness in natural and medical image segmentation tasks, including those associated with brain tumor segmentation (Akkus et al., 2017; Havaei et al., 2017; Iqbal et al., 2018; Naceur et al., 2018). However, one main issue in DCNN methods is the reliance on a large number of training data with expert annotations, which are often difficult to obtain, especially from multiple institutions. To provide such a dataset to the scientific community and a platform to compare and evaluate different automatic segmentation algorithms for brain tumors, the Multimodal Brain Tumor Segmentation Challenge (BraTS) was organized using multi-institutional pre-operative MRI scans for the segmentation of intrinsically heterogeneous brain tumor sub-regions (Menze et al., 2015; Bakas et al., 2017a,b,b), with the dataset growing every year. In the 2018 challenge, 285 training cases, 66 validation cases, and 191 testing cases were provided. Not surprisingly, DCNN-based models have quickly become the mainstream in BraTS challenges (Bakas et al., 2018). Similar to classification networks, one common DCNN method for segmentation is to use the extracted small patches to predict the class for the center voxel and slide these patches to cover the entire volume; to improve the classification accuracy of the center voxel, multi-scale patches with different receptive field sizes can be extracted simultaneously as in Kamnisas et al. (2017). In contrast, U-Net is a widely used network structure that consists of a contracting path to capture context and a symmetric expanding path that enables precise localization and segmentation for the entire input image (Ronneberger et al., 2015). If the input images and the corresponding output label maps are 3-dimensional (3D), the original U-Net construction can be extended by replacing 2D operations with their 3D counterparts (Cicek et al., 2016). However, in such cases the requirement for memory and computation speed is greatly increased so that it may not be possible to use the entire 3D volume as the input and output. To address this issue, one method is to extract smaller 3D patches as the network input and generate the label maps corresponding to these patches (Li et al., 2018). To achieve a good segmentation performance, data augmentation and optimization of patch extraction strategy and network hyper-parameters are often performed. However, in practice, it is very challenging to achieve a single "optimized" model and it is possible that any model can suffer from random errors. Using a similar concept as in traditional machine learning tasks, an ensemble of multiple models can generally improve the classification/segmentation accuracy as individual models may make different errors and by averaging or majority voting, the final number of errors can be reduced (Tan and Gilber, 2003). In this study we propose the use of an ensemble of 3D U-Nets with different hyper-parameters for brain tumor segmentation. For each 3D U-Net, the smaller 3D patches will be extracted to minimize memory overhead. To avoid extracting too many background patches and not learning sufficient information to segment tumors, a customized probability function is used to guide the patch extraction process. Furthermore, during testing, a sliding window approach is used to predict class labels with overlap between patches as a testing augmentation method to improve accuracy. On the network structure, although many new methods have been proposed that show superior performance than the U-Net in segmentation tasks, such as the densely connected network (Dense-Net) (Jegou et al., 2016; Stawiaski, 2019), a recent paper claimed that optimization on various training and testing details based on vanilla U-Net can yield robust and superior performance (Isensee et al., 2018). In our study we will compare the U-Net with Dense-Net for this task when other strategies are kept the same.

Survival prediction has a very high clinical value in prognosis and patient management. In the BraTS challenge, to demonstrate one potential clinical application of the segmentation results, the task to predict patient overall survival measured in days was also included. Additional data including patient age and resection status was provided. For training cases, the overall survival was also available for part of the dataset. Although complicated models such as DCNN or random forests (Tustison et al., 2015) can be used to capture sophisticated relationships between the input features and the output of overall survival, one main issue with these methods is overfitting, especially in this task as the training data is very small compared with the huge number of possible input features. Furthermore, the radiomics features are often difficult to explain as they lack direct clinical correspondence. Using the segmentation method proposed in this study, the sub-regions of brain tumor are expected to be accurately segmented so that various quantitative features can be calculated. To reduce overfitting, we will utilize the quantitative results and a robust linear model while limiting the number of extracted features. The correlations of these features with overall survival will also be analyzed.

METHODS

For the brain tumor segmentation task, the steps in our proposed method include pre-processing of the images, patch extraction, training multiple models using a generic 3D U-Net structure with different hyper-parameters, deployment of each model for the full volume prediction and the final ensemble step. For the survival prediction task, the steps include feature extraction, model fitting, and deployment. Data description and methodological details are provided in the following sections.

Dataset and Image Pre-processing

The datasets used in this study are provided by the BraTS challenge organizers and contains multiple-institutional clinically-acquired pre-operative multimodal MRI scans of glioblastoma (GBM/HGG) and low-grade glioma (LGG) containing (a) native (T1) and (b) post-contrast T1-weighted (T1Gd), (c) T2-weighted (T2), and (d) Fluid Attenuated Inversion Recovery (FLAIR) volumes. They were acquired with different clinical protocols and various scanners. All the imaging datasets have been segmented manually, by one to four raters, following the same annotation protocol, and their annotations were approved by experienced neuro-radiologists. Annotations comprise the GD-enhancing tumor (ET—label 4), the peritumoral edema (ED—label 2), and the necrotic and non-enhancing tumor core (NCR/NET—label 1). During training, 285 imaging cases with annotations were provided to all challenge participants. An additional 66 cases were used as validation data which did not include ground truth labels. Additionally, participants were able to upload their predictions multiple times and get the corresponding evaluation results. During the testing phase, 191 cases were provided and the teams could only upload their results once in a 48-h period and receive the final score.

To accommodate for the differences in imaging protocols, pre-processing was performed by the challenge organizers. The images from different MR sequences of the same subject were first co-registered to the same anatomical template, the SRI24 multichannel atlas of normal adult human brain (Rohlfing et al., 2010), followed by interpolation and zero-padding to the same resolution (1 mm^3) and same matrix size (240x240x155). The field-of-view (FOV) was then unified accordingly (240 mm along the left-right and anterior-posterior directions and 155 mm along the superior-inferior direction). Brain extraction was also performed using the method described in Bauer et al. (2012). To improve the homogeneity and suppress noise, N4 bias-correction (Tustison et al., 2010) and denoising using non-local means (Manjon et al., 2010) are often used in various studies. However, although these pre-processing steps can yield visually improved image quality, as shown in our previous study (Feng et al., 2018), we did not achieve an improved segmentation result on the validation data set. Considering the bias-correction and denoising algorithms are computationally intensive and time-consuming, we did not perform these two steps. To unify the intensity range, each MR sequence is scaled to be between 0 and 1.

To achieve the second task to predict patient overall survival, during training, 163 cases out of the total 285 had age, resection status and survival information available. However, the cases from The Cancer Imaging Archive (TCIA) and a few other cases did not have the resection status available so they were labeled as "NA." For all other cases, the status was either Gross Total Resection (GTR) or Subtotal Resection (STR). The survival time was given in days. During validation, 53 cases with age and resection status were provided. Similar with the segmentation task, the participants could upload the prediction multiple times. However, only 28 cases with resection status GTR were evaluated. During testing, 130 cases were provided and 77 were evaluated.

Non-uniform Patch Extraction

For simplicity, we will use foreground to denote all tumor pixels and background to denote the rest. There are several challenges in directly using the whole image as the input to a 3D U-Net: (1) the memory of a moderate GPU is often 12 Gb so that in order to fit the model into the GPU, the network needs to greatly reduce the number of features and/or the layers, which often leads to a significant drop in performance as the expressiveness of the network is much reduced; (2) the training time will be greatly prolonged since more voxels contribute to calculation of the gradients at each step and the number of steps cannot be proportionally reduced during optimization; (3) as the background voxels dominate the whole image, the class imbalance will cause the model to focus on background if trained with uniform loss, or prone to false positives if trained with weighted loss that favors the foreground voxels. Therefore, to more effectively utilize the training data, smaller patches were extracted from each subject. As the foreground labels contain much more variability and are the main targets to segment, more patches from the foreground voxels should be extracted.

In implementation, during each epoch of the training process, a random patch was extracted from each subject using non-uniform probabilities. In extraction, the voxel was first chosen as the center of the patch and the corresponding patch was extracted based on the desired size. To make sure that each extracted patch is within the whole image so that no padding is required, the voxels close to the edge of the image were excluded when determining the patch center. From all voxels valid to be the patch center, the sampling was performed based on the probability function $p_{i,j,k}$ calculated using the following equation:

$$p_{i,j,k} = \frac{s_{i,j,k}}{\sum_{i,j,k} s_{i,j,k}} \tag{1}$$

in which $s_{i,j,k} = 1$ for all voxels with maximal intensity lower than the 1st percentile, $s_{i,j,k} = 6$ for all foreground voxels and $s_{i,j,k} = 3$ for the rest. These values were picked to greatly favor the tumor regions and slightly favor the regions with normal brain tissue compared with the background voxels. However, the exact ratio was determined empirically without rigorous tuning. For each training iteration, one patch was extracted using this method. Since normal brain images are symmetric along the left-right direction, a random flip along this direction was made after patch extraction. No other augmentation was applied.

Before training, the per-input-channel mean and standard deviation of extracted patches were calculated by running the extraction process 400 times, with each time using a randomly selected training subject. The extracted patches were then normalized by subtracting the mean and dividing by the standard deviation along each input channel.

Network Structure and Training

A 3D U-Net based network was used as the general structure, as shown in **Figure 1**. Zero padding was used to make sure the spatial dimension of the output is the same with the input. For each encoding block, a VGG like network (Simonyan and Zisserman, 2014) with two consecutive 3D convolutional layers

of kernel size 3 followed by the activation function and batch norm layers were used. The parametric rectilinear function (PReLU) (Xu et al., 2015), given as:

$$f(x) = \max(0, x) - \alpha \max(0, -x) \quad (2)$$

was used as the activation function (with trainable parameter α). The number of features was doubled while the spatial dimension was halved with every encoding block, as in the conventional U-Net structure. A dropout layer with ratio 0.5 was added after the last encoding block. Symmetric decoding blocks were used with skip-connections from corresponding encoding blocks. Features were concatenated to the de-convolution outputs. The extracted segmentation map of the input patch was expanded to the multi-class the ground truth labels (3 foreground classes and the background). Cross entropy was used as the loss function. In addition to a uniform loss among all classification labels, the weighted loss, in which different labels can be assigned with different weights, was also used.

It is shown that a wider network with large number of features and a deeper network can increase the expressiveness and thus performance of the network (Wu et al., 2016); furthermore, the larger the patch size, the more spatial information to be used in one patch; however, as mentioned before, the memory of the GPU is often a limiting factor with 3D inputs. In our study, we balanced the three parameters (number of encoding/decoding blocks, input features at the first layer and patch size) to make sure that the GPU memory is sufficient while favoring one in one model. Specifically, if the patch size is increased, to keep the same rule of doubling the number of filters every block, the number of blocks cannot be more than 3 without exceeding GPU memory. The exact choice of these parameters was made empirically with the general principle to be as different as possible to reduce the correlations of random errors by a single parameter set. In addition, the weighted loss function, which favors the foreground voxels, can often improve the sensitivity but sacrifice specificity as it punishes more for missed foreground segmentations. Therefore, for each combination of these parameters, we used both the weighted and uniform loss functions. Although the increase of the number of models may further benefit the final results, in a way that is similar with more averages, the time for training and testing will also increase proportionally. Therefore, a total of six model was selected, with detailed parameters shown in **Table 1**. N denotes the input patch size, M denotes the number of encoding/decoding blocks and f denotes the input features at the first layer. For the weighted loss function, 1.0 was used for background and 2.0 was used for each of the foreground classes.

TABLE 1 | Detailed parameters for all 6 3D U-Net models.

Model#	M	N	f	Loss Type
1	3	64	96	Uniform
2	3	64	96	Weighted
3	4	64	96	Uniform
4	4	64	96	Weighted
5	3	80	64	Uniform
6	3	80	64	Weighted

Training was performed on a Nvidia Titan Xp GPU with 12 Gb memory. Six hundred forty epochs were used. As mentioned earlier, during each epoch, only one patch was extracted from every subject. Subject orders were randomly permuted every epoch. Implementation was based on the TensorFlow framework. Batch size was set to 1 during training. During testing, due to the sensitivity associated with smaller batch sizes, all batch norm layers did not use the running statistics but the statistics of the batch itself. This is the same as instance normalization (Ulyanov et al., 2016) when the batch size is 1 as it normalizes each feature map with its own mean and standard deviation. The Adam optimizer was used with an initial learning rate of 0.0005 without further adjustments during training as it can self-adjust the rate of gradient update so that no manual reduction of learning rate is necessary (Kingma and Ba, 2014). The total training time was about 60 h.

Deployment of Each Segmentation Model and Ensemble

Although the fully convolutional segmentation network can be applied to the input images of any size, due to the fact that the whole network with the entire image as the input cannot fit into the memory during deployment, a sliding window approach needs to be used to get the output for each subject. However, as significant padding was made to generate the output label map at the same size as the input, boundary voxels of a patch were expected to yield unstable predictions when sliding the window across the whole image without overlaps. To alleviate this problem, a stride size at a fraction of the window size was used and the output probability was averaged. In implementation, the deployment window size was chosen to be the same as the training window size, and the stride was chosen as ½ of the window size. For each window, the original image and left-right flipped image were both predicted, and the average probability after flipping back the output of the flipped input was used as the output. Therefore, each voxel, except for a few on the edge, will be predicted 16 times when sliding across all directions. Although smaller stride sizes can be used to further improve the accuracy with more averages, the deployment time will be increased 8 times for every ½ reduction of the window size and thus quickly becomes unmanageable. Using the parameters as mentioned on the same GPU, it took about 1 min to generate the output for the entire volume per subject. Instead of performing a thresholding on the probability output to get the final labels, the direct probability output after the last convolutional layer was saved for each model as a measure of "confidence" for each model.

The ensemble modeling process was rather straightforward. The probability output of all classes from each model was averaged to get the final probability output. The class with the highest probability was selected as the final segmentation label for each voxel.

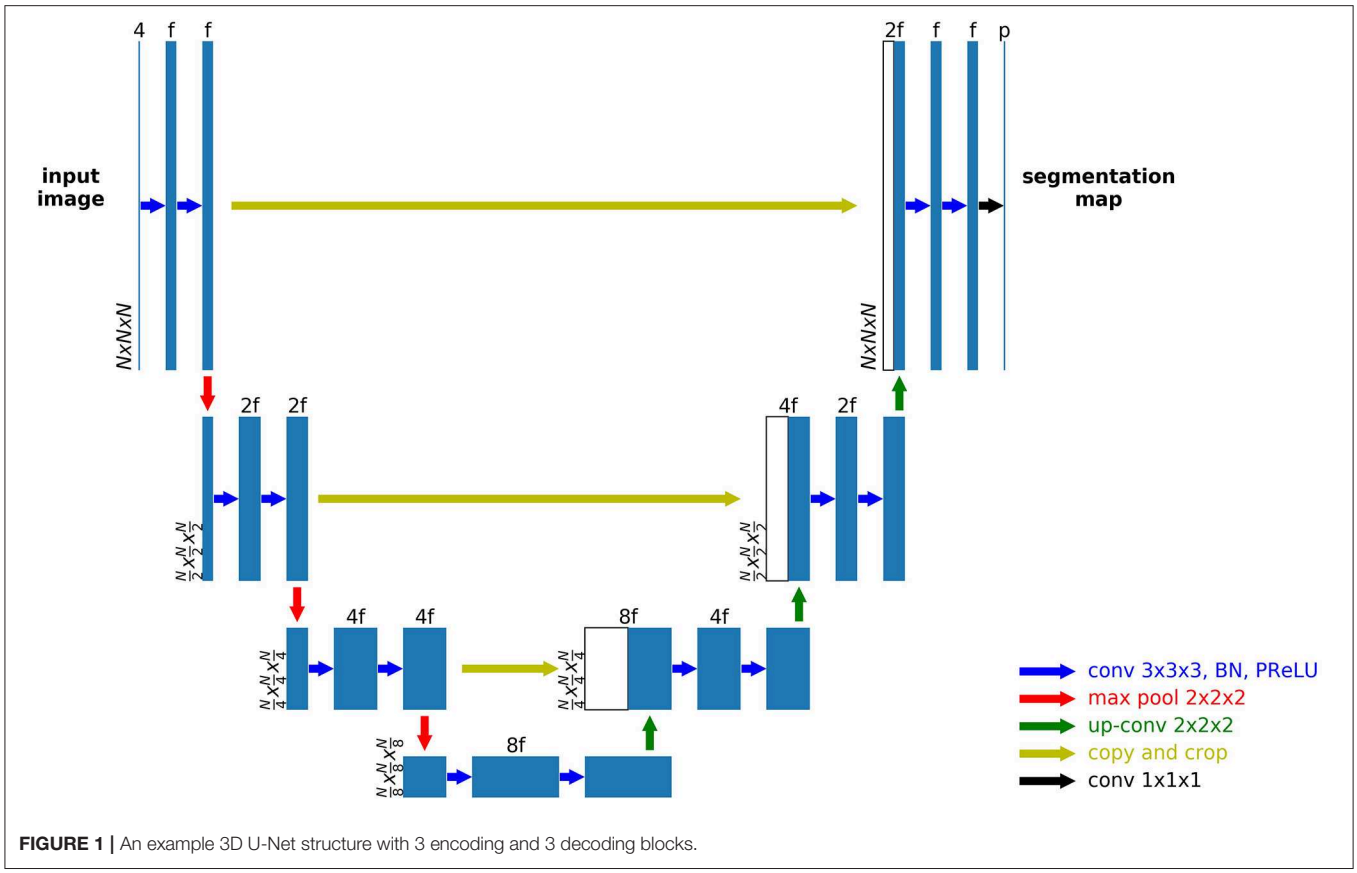


FIGURE 1 | An example 3D U-Net structure with 3 encoding and 3 decoding blocks.

Comparison of U-Net and Dense-Net

The Dense-Net was implemented following the standard structure as in Jegou et al. (2016). Specifically, the block number was 4, layers per block was 12 and the growth rate was 12. In terms of architecture, the Dense-Net-BC (further compression) was used. The uniform cross entropy function was used as the loss function. As a fair comparison, only the U-Net with 4 encoding/decoding blocks and uniform loss function (model 3 in **Table 1**) was compared. The patch extraction and augmentation were kept the same for the two models. As the evaluation using the BraTS validation and testing datasets requires submission to the server of the BraTS organizers, which has a limit on the number of allowed submissions, we only used the BraTS training dataset and randomly split it with a 3:2 ratio for training and validation in this comparison experiment.

Survival Prediction

To predict the post-surgery survival time measured in days, extracted imaging features and non-imaging features were used to construct a linear regression model. As MR images often exhibit variations in imaging intensity and contrast, the intensity values of the images were not directly used in our survival modeling. Instead, six simple volumetric features were calculated from the segmented labels of the three tumor sub-regions: the enhancing tumor core, non-enhancing and necrotic region and edema, with two features per region. During training, the ground truth label maps were used; during validation and testing, the automatically segmented label maps were used. For each foreground class, the volume (V) was determined by summing up the voxels whereas the surface area (S) was calculated by summing up the magnitude of the gradients along three directions, as described in the following equations

$$V_{ROI} = \sum_{i,j,k} s_{i,j,k} \tag{3}$$

$$S_{ROI} = \sum_{i,j,k} s_{i,j,k} \sqrt{\left(\frac{\partial s}{\partial i}\right)^2 + \left(\frac{\partial s}{\partial j}\right)^2 + \left(\frac{\partial s}{\partial k}\right)^2} \tag{4}$$

in which ROI denotes a specific foreground class and $s_{i,j,k} = 1$ for voxels that are classified to belong to this ROI and $s_{i,j,k} = 0$ otherwise. The volume represents the size of each sub-region and thus may reflect the severity of the tumor. It is expected that the larger the volume, the poorer the prognosis. The surface area is another measure of the size; however, together with volume, it can also serve as a measure for the shape. Given a fixed volume, the more irregular the shape, the larger the surface area; therefore, a larger surface area may indicate the aggressiveness of the tumor and the increased difficulty in surgery.

Age and resection status were used as non-imaging clinical features. As there were two classes of resection status and many missing values of this status, a two-dimensional feature vector

was used to represent the status, given as GTR: (1, 0), STR: (0, 1), and NA: (0, 0). A linear regression model was employed after normalizing each input feature to zero mean and unit standard deviation. As the input feature size is 9, the risk for overfitting is greatly reduced.

For evaluation, in addition to mean and median square error of survival time predictions, the classification of subjects as long-survivors (e.g., >15 months), short-survivors (e.g., <10 months), and mid-survivors (e.g., between 10 and 15 months) was performed. For the challenge, ranking of the participating teams was based on accuracy (i.e., the number of correctly classified patients) with respect to this grouping.

RESULTS

Comparison of U-Net and Dense-Net

Among the 285 training subjects, 171 were used for training the two models and 114 were used for testing. The dice indexes of the enhanced tumor (ET), whole tumor (WT) and tumor core (TC) were calculated and compared, as shown in **Figure 2**. The blue bars show the results from U-Net and the green bars show those from Dense-Net. The two methods yield very similar performances with the Dense-Net having slightly better performance in tumor core. However, the paired Student's *t*-test was performed between the two methods and showed no statistically significant differences when the threshold of *p*-value was set at 0.05.

Brain Tumor Segmentation

All 285 training subjects were used in the training process. 66 subjects were provided as validation. The dice indexes, sensitivities and specificities, 95% Hausdorff distances of ET, WT, and TC were automatically calculated after submitting to the CBICA's Image Processing Portal. ET corresponds to label 4 in the direct output label maps; WT is the union of all non-background label maps including label 1, 2, and 4; TC is the union of ET and NCR/NET, or label 1 and 4. With multiple submissions, we were able to compare the performances of each individual model and the final ensemble.

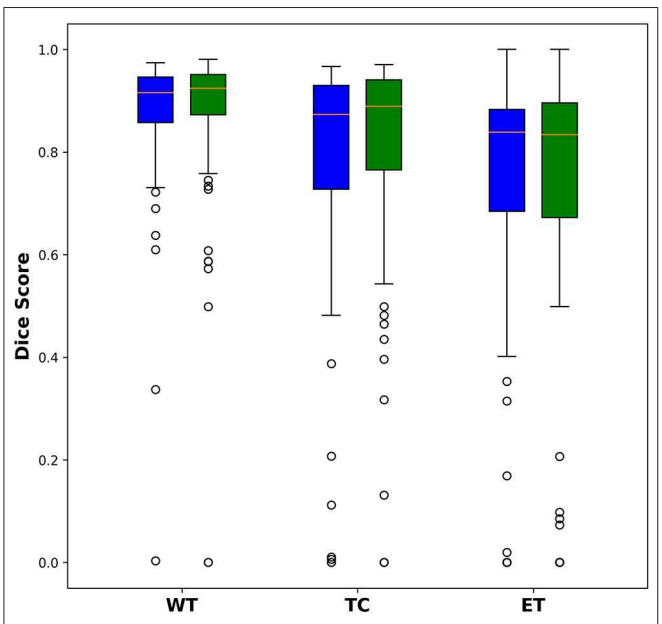


FIGURE 2 | Comparison of dice indexes using U-Net and Dense-Net. Green bars show the results using U-Net; blue bars show the results using Dense-Net. The two models have very similar performances without any statistically significant differences.

Table 2 shows the mean Dice scores (Dice) and 95% Hausdorff distances (Dist) of ET, WT and TC in mm for the 6 individual models and the ensemble of them. The model with the best performance of each metric is highlighted. For the evaluation, sensitivity and specificity were also calculated to determine over- or under-segmentations of tumor sub-regions. Detailed descriptions of the evaluation metrics were provided in Menze et al. (2015). As we found that sensitivity and specificity were highly correlated with the Dice indexes, they are not included in the table. The best performance of each evaluation metric is highlighted. For WT, all 3D U-Net models perform similarly, except for a slightly worse performance with model 4. However, model 4 has the highest Dice for ET. The rankings based on Dice scores are also not consistent with the rankings based on the distance measures. This shows that no single parameter set has clear advantage over others. However, the ensemble of them has the best overall Dice scores as compared with each individual model. Paired student's *t*-tests were performed between each model and the ensemble on Dice scores with the red scores showing statistically inferior performances of one model compared with the ensemble ($p < 0.05$). For WT, model 1–5 all showed significant inferior performances. For model 6, although no statistical significance was found, the *p*-values were close to 0.05. The distance metrics show a wider range and the ensemble does not achieve the smallest values. However, as the Haudorff distance is largely determined by the "worst" pixels, it may be less reliable in obtaining an overall performance evaluation as compared with Dice scores. Despite this, the metrics in the ensemble method for all three sub-regions are all on the lower end, showing increased robustness. It is also noticed that weighted cross-entropy loss has high sensitivity but lower specificity compared with the uniform counterpart, which is likely due to the fact that by assigning more weights to the foreground, the network tends to be more aggressive in assigning foreground labels.

Figures 3, **4** show two slices (axial slice 76 and 81) of the automatically segmented labels overlaid on the

TABLE 2 | Performances of each individual model and the ensemble.

Model #	Dice_ET	Dice_WT	Dice_TC	Dist_ET	Dist_WT	Dist_TC
1	0.7839	0.9061	0.8233	4.0496	4.0401	6.5389
2	0.7681	0.9070	0.8126	4.2215	6.1359	**6.0561**
3	0.7538	0.9072	0.8236	4.7615	5.7021	9.0000
4	0.7874	0.9001	0.8088	**3.9195**	6.3093	6.9586
5	0.7704	0.9061	0.8227	4.0314	4.7068	6.5905
6	0.7819	0.9097	0.8217	3.9368	**3.6666**	6.3705
Ensemble	**0.7946**	**0.9114**	**0.8304**	3.9679	3.7842	6.5234

Bold values show the model with the best performance of each metric.

T1Gd and T2 images, respectively. The showed case was "Brats18_CBICA_BHF_1" and was randomly selected from the validation dataset for demonstration. A single model may suffer from under- or over-segmentation while the average of multiple models achieves a more stable performance, which is also closer to the ground-truth, as shown with the improved Dice scores. Furthermore, the ensemble of all 6 models yields a much smoother boundary for different sub-regions and eliminates a few isolated regions, which are likely false positives.

In 191 testing cases, as only one submission was allowed, we submitted the final ensemble results for evaluation. The mean Dice scores for ET, WT and TC were 0.754, 0.878, and 0.799 and the 95% Hausdorff distances were 20.29, 7.41, and 22.06 mm, respectively. It is also noted that 2 of the testing cases failed to predict any tumor voxels, resulting in Dice scores of 0. Compared with validation cases, the average performance for testing cases was much worse.

The paper published by the challenge organizers (Bakas et al., 2018) summarized the performance by all 63 participating teams, including ours. The ranking was based on the testing cases as only one submission is allowed to avoid learning from the submissions. Our team (xfeng) achieved the 9th place in the segmentation task [Figure 7 in Bakas et al. (2018)]. However, the differences among the top teams were relatively small.

Overall Survival Prediction

Figure 5 plots the extracted features against the overall survival in the training data. The correlation coefficients between the six radiomic features from images and the overall survival were also calculated as well as between age and the survival. Negative correlations between imaging features and the survival are observed, indicating that the larger the specific tumor sub-regions, the shorter the survival will be. To better illustrate this trend, we binned the overall survival into the short-term ($<$10 months), medium-term (10–15 months) and long-term ($>$15 months) and drawn the box plots for survivals, as shown in **Figure 6**. The general trend is consistent with the previous results, showing that the larger the volume and surface, the worse the prognosis. The correlation between age and survival is also expected. Furthermore, the correlation between age and survival is the strongest among all selected features. For resection status, patients who underwent GTR have longer survival rates than the STR patients. However, no statistical differences were found using a Student's *t*-test.

A multivariate linear regression model was trained with all the features from 163 training subjects. For the 28 validation cases, the accuracy was 0.321. The mean and median errors were 314.8 and 278.85 days, respectively. For the 77 testing cases, accuracy was 0.61 corresponding to mean and median errors of 481.4 and 185.22 days, respectively. It should be noted that the accuracy for the testing cases was much higher than for the validation cases. We did not use the validation cases to tune any parameters in training the model due to potential overfitting. Our testing performance ranked 1st among all participants, indicating the robustness of the linear model. Compared with other participating teams, who used radiomics and/or machine learning based modeling, this simple strategy yielded the best accuracy. It is noted that one team used the age as the only predictor and used a linear regression model similar to our method and achieved the 3rd place in survival task, as summarized in Bakas et al. (2018).

DISCUSSION AND CONCLUSIONS

In this paper we developed a brain tumor segmentation method using an ensemble of 3D U-Nets. Six networks with different numbers of encoding/decoding blocks, input patch sizes and different weights for loss were trained and ensembled together by averaging the final prediction probabilities. The results showed improvements with the ensemble model compared with any of the single models. For the survival prediction task, we extracted six simple features from the segmentation labels and used a multivariate linear regression by combining them with non-imaging clinical features such as age and resection status. The survival prediction achieved 1st place among all challenge participants.

In terms of network structure, we found it very difficult to pick the "best" model and/or hyper-parameter set since most models perform very similarly. The comparison between U-Net and Dense-Net showed that it is hard to pick a clear winner for network structure. It is indeed one disadvantage of DCNN as the "black-box" nature of the network makes it challenging to analyze the effect of network structure and parameter except from the final performance. Furthermore, the extremely long computation times and randomness in training the model and selected validation datasets makes comparison of different models difficult. In this paper, we empirically determined a few design options such as the usage of 3D U-Net and non-uniform patch extraction. Multiple models with architectural variation can form an ensemble to overcome random errors made by any individual model. Similar to using averages in measurements to improve signal-to-noise ratio, in which the marginal increase of performance can reduce as the number of averages increases, we aim to strike a balance between training and validation time and the expected performance. The ensemble yielded an improved performance in both quantitative measures and visual examination; however, one limitation of our approach is the lack of objective measures to achieve optimal combination of models. Instead, we empirically determined number of models to be 6 and chose the corresponding hyperparameters. An interesting alternative is to use grid search to gain an optimal set of hyperparameters, which is currently a popular research topic; however, one possible concern is that this may lead to overfitting as the validation set is much smaller (66 cases) compared with the training and testing dataset; to mitigate this concern, N-fold cross validation can be used in combination with the grid search method, which will be performed in future studies.

Compared with the patch-based model that only predicts the center pixel, the 3D U-Net predicts the segmentation label map for the full input. As it is limited by the GPU memory to use the full image as the input, smaller patches are extracted. However, this can lead to reduced receptive field,

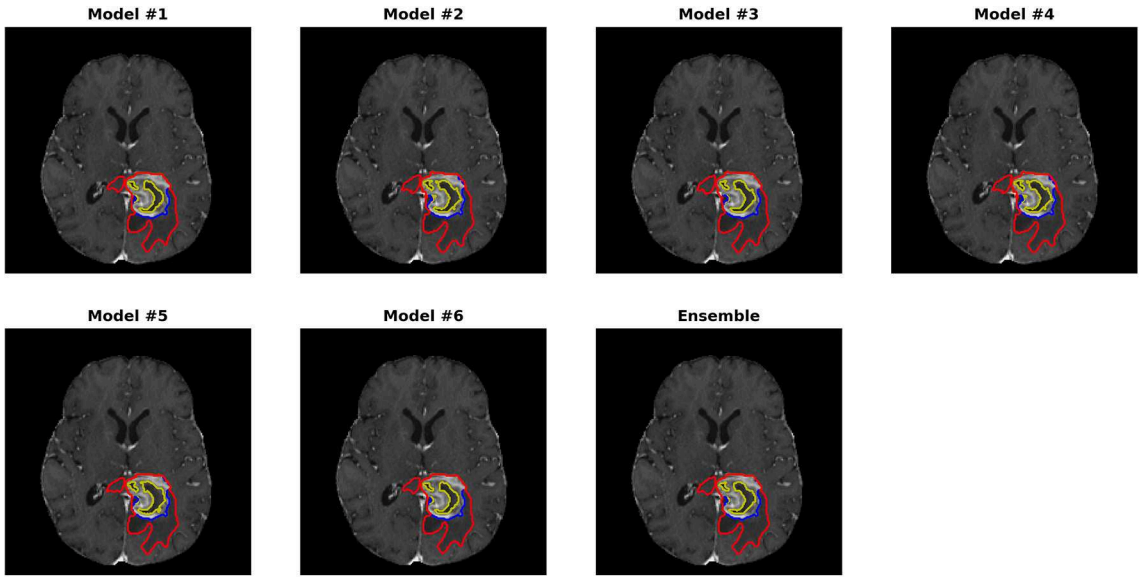


FIGURE 3 | Automatically segmented sub-regions from models 1–6 and the ensemble model. The underlying image is the corresponding T1Gd from the validation case "Brats18_CBICA_BHF_1." Red, yellow and blue delineate the predicted boundaries of the total tumor, enhanced tumor core, and peritumoral edema, respectively.

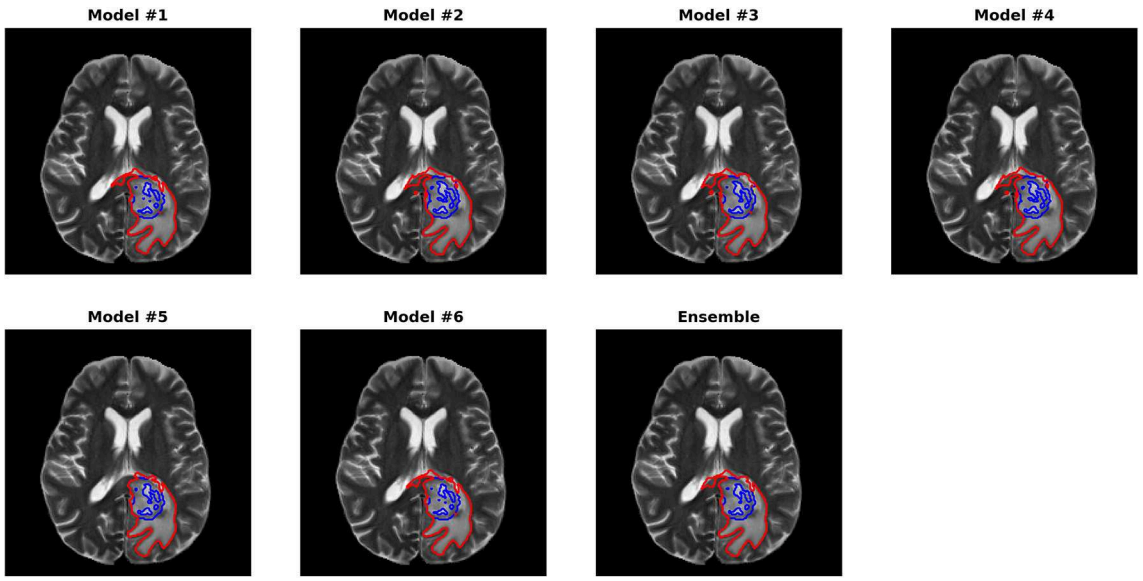


FIGURE 4 | Automatically segmented sub-regions from model 1–6 and the ensembled model. The underlying image is the corresponding T2 from the validation case "Brats18_CBICA_BHF_1." Red, yellow and blue delineate the predicted boundaries of the total tumor, enhanced tumor core, and peritumoral edema, respectively.

which is even worse for the pixels on the edge as only half of the receptive field contains information. We hypothesize that with a much larger patch size such as 128x128x128, the performance can be improved, however, the majority of GPUs only have 12 Gb memory, which cannot deal with such an input without significantly sacrificing the network complexity. To overcome the reduced receptive field of the edge pixels, we used significant overlap during deployment in a sliding windows fashion and average the output, which shows performance improvement.

For pre-processing, although in many studies the bias correction was commonly used, as in our previous experiment, we did not find any significant benefit in the proposed method. Although bias correction can greatly improve the

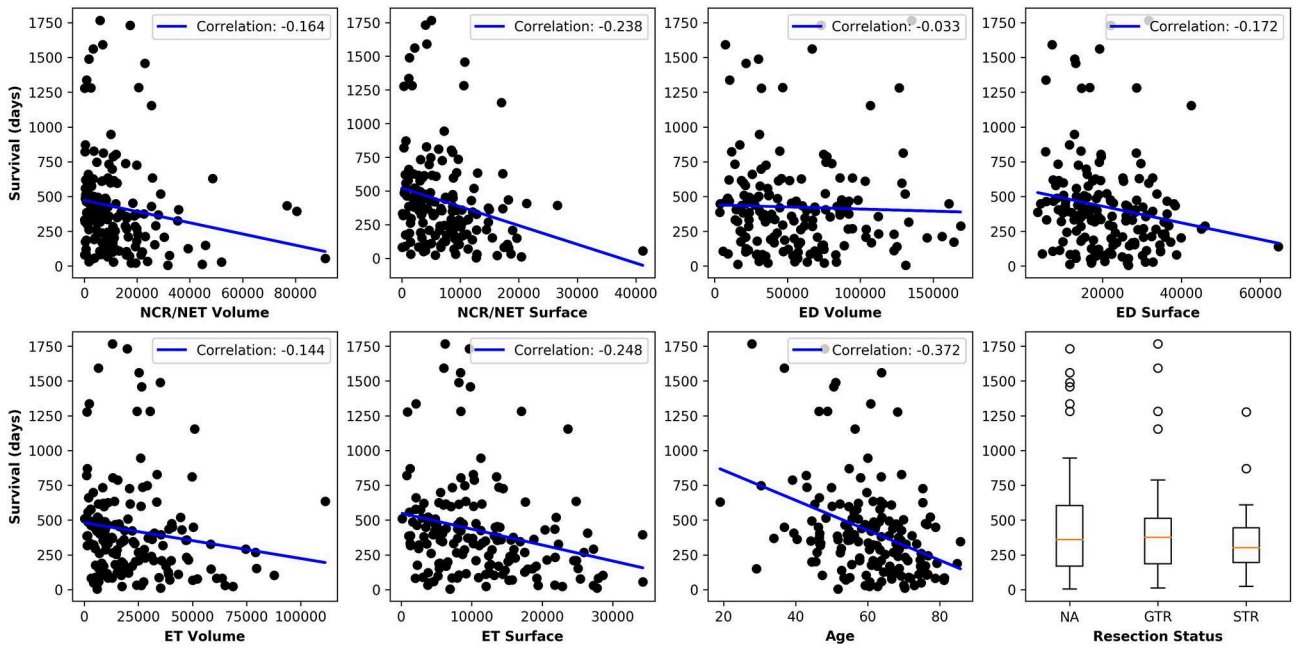


FIGURE 5 | Relationships between each extracted feature and the overall survival. The correlation coefficients are shown as well. Age shows the strongest negative correlation with survival and all imaging features show moderate to very weak negative correlations.

quality of the image by removing inhomogeneity artifacts and thus segmentation performance for any intensity-based method, DCNN may be able to learn and overcome any bias in the image so that it may not be necessary to pre-compensate for it. As it is time-consuming to run the bias correction, we did not perform this step in our final experiments. However, additional experiments on other datasets are required to continue the investigation on this topic.

For the segmentation results, as we can get the evaluation for each individual case, it is noted that the median metrics were significantly higher than the mean metrics. For example, the median Dice scores were 0.870, 0.926, and 0.911 for ET, WT, and TC in the final ensemble model. It makes sense in that the theoretical maximum Dice score is 1 and minimum Dice score is 0. However, we noted that in several cases, the Dice scores were as low as 0 for ET and TC, meaning that the model completely missed the corresponding regions. **Figure 7** shows an example ("Brats18_TCIA10_195_1") with 0 Dice score for ET with post-contrast T1. Red and blue show the contours for WT and edema, respectively. No ET is detected in this case. However, it is indeed very difficult to identify the enhanced regions as the contrast enhancement is weak. Most of the subjects had the WT Dice score larger than 0.9, indicating very high segmentation quality. However, one case had a much lower Dice of 0.63. A careful examination showed that this case predicted a very small tumor region and the contrast was visually weak. It shows that although in most cases the automatic segmentation yields very accurate result, in difficult cases with reduced contrast and/or small tumor region, the automatic result may be sub-optimal and manual expert examination and correction is still required.

Comparing the testing with validation cases, we noticed a significant gap in performance. Due to the design of the challenge, the participants can submit multiple times for the validation cases to gain any performance improvements so that the model may overfit on the validation cases; however, in our study we did not use the validation cases to perform any hyper-parameter tuning to select an optimal model. Therefore, the performance differences are likely due to more difficult cases in the testing dataset, including the two that the model completely failed. One possible reason is that the testing data covers a wide range of MR imaging protocols and field strength, some even with moderate to severe artifacts due to motion and/or inhomogeneity in one or multiple sequences, causing difficulty in achieving a consistent segmentation performance. Further investigation to continue to improve the performance and robustness of the model, especially for these difficulty cases, will be performed.

Our segmentation method ranked 9th in the challenge. The 1st place winner used a patch size of 128x128x128 with autoencoder regularization (Myronenko, 2019) and the 2nd place used an optimized U-Net (Isensee et al., 2019). As all the top teams had very similar performances and there were many different detailed strategies in implementation, it is unclear which ones are the dominating factor for the superior performance. One possible strategy is to apply post-processing to our method as the removal of vessels may have a significant impact on the final score. We also participated in the 2017 BraTS challenge using a single model (model 1) and ranked 6th in it [Figure 5 in Myronenko (2019)], showing that the U-Net can be competitive in this challenge with optimization. Further study will be performed on this topic.

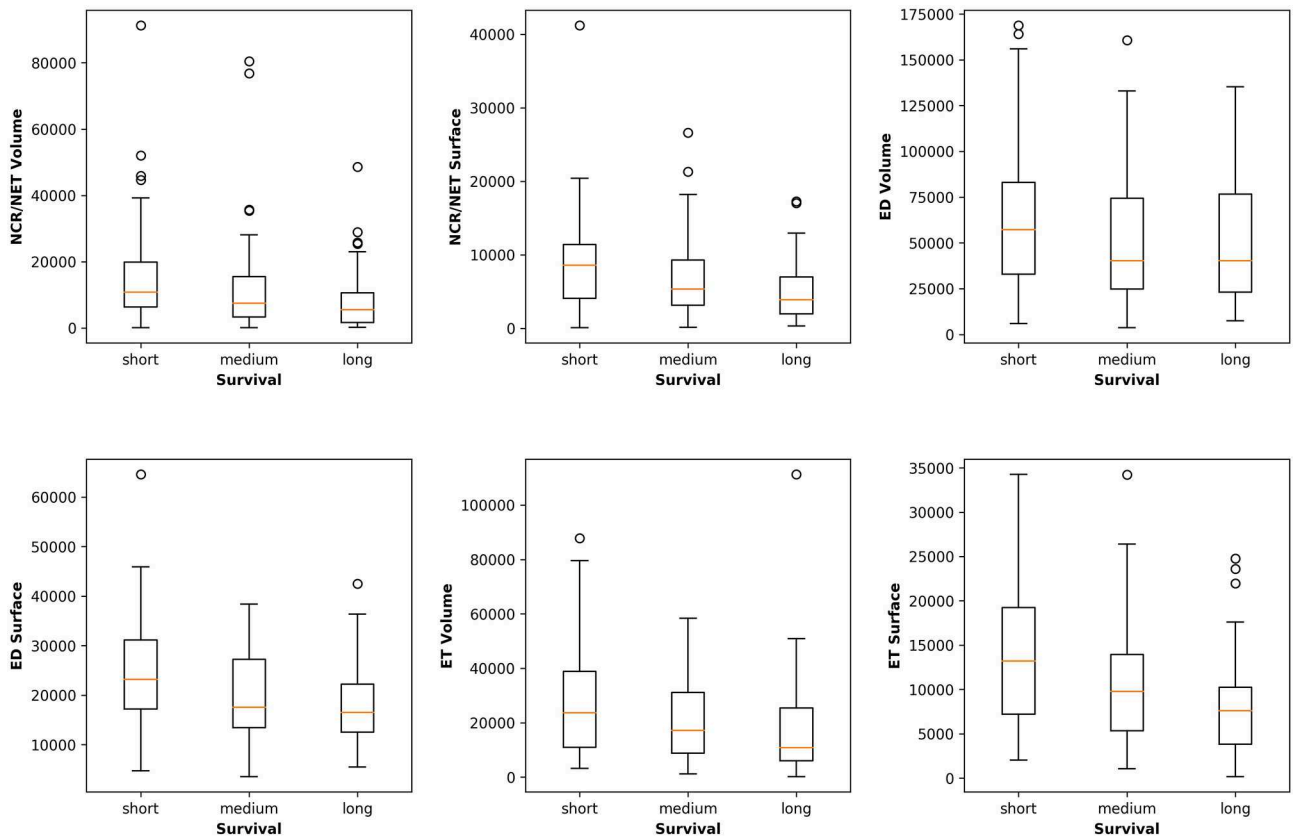


FIGURE 6 | Relationships between the volume and surface features and the binned overall survival (short-, medium-, and long-term survival). The general trend shows that the smaller the volume and surface, the longer the survival.

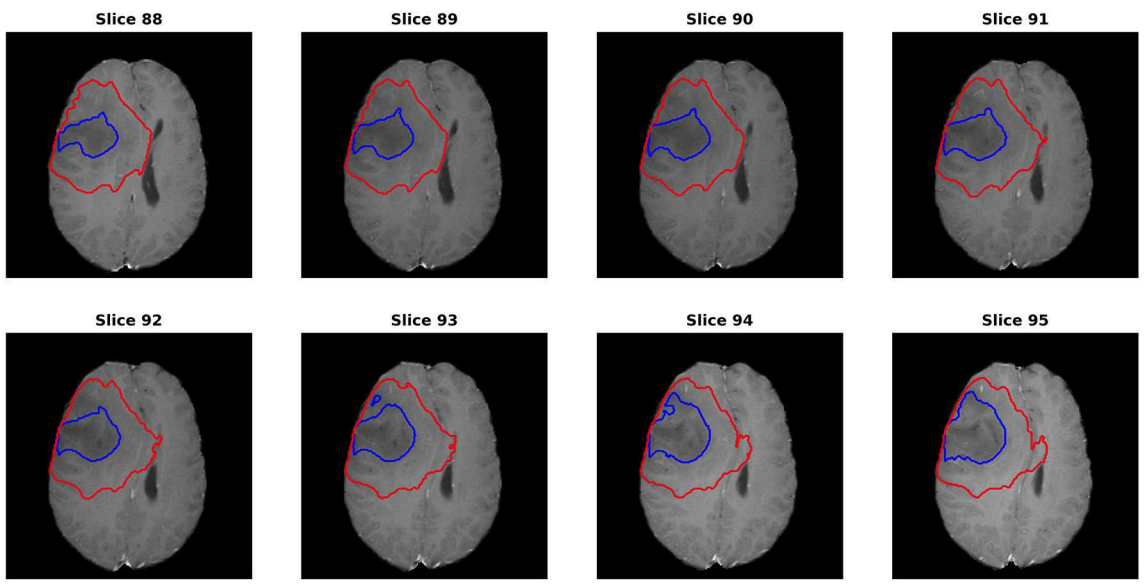


FIGURE 7 | One case ("Brats18_TCIA10_195_1") from the validation dataset showing failed segmentation for ET region on post-contrast T1. Red and blue show the boundaries for whole tumor and edema, respectively. No enhanced region is detected. It is indeed very difficult to determine the exact enhanced areas in the images, even for human experts.

For the survival prediction task, since the model is very likely to overfit with the given small dataset and since patient overall survival is affected by many aspects which are not captured in this dataset, we used a multivariate linear regression model as the safest option to minimize overfitting, although at the cost of its expressiveness. As volumetric features are assumed to be most relevant to overall survival, we only included the volumes and surface areas of different sub-regions and ignored other high-order features to reduce overfitting. In addition, these features are easy to interpret as they have direct clinical correspondences; therefore, their clinical adoption can be potentially much easier. This proved to be effective in the challenge; although exploration of additional features and more expressive models with a larger dataset could possibly improve the accuracy of survival prediction. Furthermore, adding other clinical features such as molecular and genetic types may continue to improve the accuracy of prognosis.

In conclusion, we developed an automatic brain tumor segmentation method using an ensemble of 3D U-Nets and showed the superiority over a single model. Based on the segmentation results, we extracted a few simple features and examined their correlations with the overall survival. A multivariate linear regression model was trained to predict the survival and showed high accuracy.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XF conducted all experiments and drafted the manuscript. NT, SP, and CM provided significant assistance in technical issues and writing. SP provided guidance in clinical application of this approach and CM provided guidance in the acquisition protocol using MRI.

FUNDING

University of Virginia Engineering-in-Medicine seed grant. NVIDIA GPU Grant.

ACKNOWLEDGMENTS

We acknowledge the organizers of the MICCAI Multimodal Brain Tumor Segmentation Challenge 2018 in making a successful challenge and providing opportunities for participants to develop and compare their algorithms. This manuscript has been released as a Pre-Print at https://arxiv.org/ftp/arxiv/papers/1812/1812.01049.pdf.

REFERENCES

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erikson, B. J. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *J. Digit. Imaging* 30, 449–459. doi: 10.1007/s10278-017-9983-4

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection*. The Cancer Imaging Archive.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). 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. [cs.CV].

Bauer, S., Fejes, T., and Reyes, M. (2012). A skull-stripping filter for ITK. *Insight J.* 70–78. Available online at: https://www.insight-journal.org/browse/publication/859

Cha, S. (2006). Update on brain tumor imaging: from anatomy to physiology. *AJNR Am. J. Neuroradiol.* 27, 475–487.

Cicek, O., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. *arxiv [Preprint] arXiv*:1606.06650. [cs.CV]. doi: 10.1007/978-3-319-46723-8_49

Feng, X., Tustison, N., and Meyer, C. (2018). Brain tumor segmentation using an ensemble of 3D U-nets and overall survival prediction using radiomic features. *arxiv [Preprint] arXiv*:1812.01049. [cs.CV]. doi: 10.1007/978-3-030-11726-9_25

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: images are more than pictures. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with Deep Neural Networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Iqbal, S., Ghani, M. U., Saba, T., and Rehman, A. (2018). Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN). *Microsc. Res. Tech.* 81:419–427. doi: 10.1002/jemt.22994

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2019). "No new-net," in *International MICCAI Brainlesion Workshop, BrainLes 2018: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Granada: Springer), 234–244.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., et al. (2018). nnU-net: self-adapting framework for U-net-based medical image segmentation. *arxiv [Preprint] arXiv*:1809.10486. [cs.CV]. doi: 10.1007/978-3-658-25326-4_7

Jegou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2016). The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. *arxiv [Preprint] arXiv*:1611.09326. [cs.CV]. doi: 10.1109/CVPRW.2017.156

Kamnisas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for Accurate Brain Lesion Segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arxiv [Preprint] arXiv*:1412.6980. [cs.LG].

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., et al. (2012). Radiomics: the process and the challenges. *Magn. Reson. Imaging* 30, 1234–1248. doi: 10.1016/j.mri.2012.06.010

Li, J., Sarma, K. V., Chung Ho, K., Gertych, A., Knudsen, B. S., and Arnold, C. W. (2018). A Multi-scale U-Net for semantic segmentation of histological images from radical prostatectomies. *AMIA Annu. Symp. Proc.* 16, 1140–1148.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Manjon, J. V., Coupe, P., Marti-Bonmati, L., Collins, D. L., and Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. *J. Magn. Reson. Imaging* 31, 192–203. doi: 10.1002/jmri.22003

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Myronenko, A. (2019). 3d mri brain tumor segmentation using autoencoder regularization. *BrainLes* 11384, 254–266. doi: 10.1007/978-3-030-11726-9_28

Naceur, M. B., Saouli, R., Akil, M., and Kachouri, R. (2018). Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images. *Comput. Methods Programs Biomed.* 166, 39–49. doi: 10.1016/j.cmpb.2018.09.007

Rohlfing, T., Zahr, N. M., Sullivan, E. V., and Pfefferbaum, A. (2010). The SRI24 multichannel atlas of normal adult human brain structure. *Hum. Brain Mapp.* 31, 798–819. doi: 10.1002/hbm.20906

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *arxiv [Preprint] arXiv*:1505.04597. [cs.CV]. doi: 10.1007/978-3-319-24574-4_28

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arxiv [Preprint] arXiv*:1409.1556. [cs.CV].

Stawiaski, J. (2019). A Pretrained DenseNet Encoder for Brain Tumor Segmentation. *arxiv [Preprint] arXiv*:1811.07542. [cs.CV]. doi: 10.1007/978-3-030-11726-9_10

Tan, A. C., and Gilber, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinformatics* 2(Suppl. 3), S75–S83.

Tustison, N. J., Avantes, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Tustison, N. J., Shrinidhi, K. L., Wintermark, M., Durst, C. R., Kandel, B. M., Gee, J. C., et al. (2015). Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (Simplified) with ANTsR. *Neuroinformatics* 13, 209–225. doi: 10.1007/s12021-014-9245-2

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. *arxiv [Preprint] arXiv*:1607.08022. [cs.CV].

Upadhyay, N., and Waldman, A. D. (2011). Conventional MRI evaluation of gliomas. *Br. J. Radiol.* 84, S107–S111. doi: 10.1259/bjr/65711810

Wrensch, M., Minn, Y., Chew, T., Bondy, M., and Berger, M. S. (2002). Epidemiology of primary brain tumors: current concepts and review of the literature. *Neuro Oncol.* 4, 278–299. doi: 10.1093/neuonc/4.4.278

Wu, Z., Shen, C., and van den Hengel, A. (2016). Wider or deeper: revisiting the resnet model for visual recognition. *arxiv [Preprint] arXiv*:1611.10080. [cs.CV].

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arxiv [Preprint] arXiv*:1505.00853. [cs.LG].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 08 April 2020
doi: 10.3389/fnins.2020.00282

Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation

Alain Jungo[1,2]*, *Fabian Balsiger*[1,2] *and Mauricio Reyes*[1,2]

[1] *Insel Data Science Center, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland,* [2] *ARTORG Center, University of Bern, Bern, Switzerland*

Automatic segmentation of brain tumors has the potential to enable volumetric measures and high-throughput analysis in the clinical setting. Reaching this potential seems almost achieved, considering the steady increase in segmentation accuracy. However, despite segmentation accuracy, the current methods still do not meet the robustness levels required for patient-centered clinical use. In this regard, uncertainty estimates are a promising direction to improve the robustness of automated segmentation systems. Different uncertainty estimation methods have been proposed, but little is known about their usefulness and limitations for brain tumor segmentation. In this study, we present an analysis of the most commonly used uncertainty estimation methods in regards to benefits and challenges for brain tumor segmentation. We evaluated their quality in terms of calibration, segmentation error localization, and segmentation failure detection. Our results show that the uncertainty methods are typically well-calibrated when evaluated at the dataset level. Evaluated at the subject level, we found notable miscalibrations and limited segmentation error localization (e.g., for correcting segmentations), which hinder the direct use of the voxel-wise uncertainties. Nevertheless, voxel-wise uncertainty showed value to detect failed segmentations when uncertainty estimates are aggregated at the subject level. Therefore, we suggest a careful usage of voxel-wise uncertainty measures and highlight the importance of developing solutions that address the subject-level requirements on calibration and segmentation error localization.

Keywords: segmentation, brain tumor, uncertainty estimation, quality, deep learning

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Suyash P. Awate,
Indian Institute of Technology Bombay, India
Roberto Viviani,
University of Innsbruck, Austria
Tal Arbel,
McGill University, Canada

***Correspondence:**
Alain Jungo
alain.jungo@artorg.unibe.ch

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *30 September 2019*
Accepted: *12 March 2020*
Published: *08 April 2020*

Citation:
Jungo A, Balsiger F and Reyes M (2020) Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. Front. Neurosci. 14:282. doi: 10.3389/fnins.2020.00282

1. INTRODUCTION

Automated segmentation holds promise to improve the treatment of brain tumors by providing more reliable volumetric measures for treatment response assessment (Reuter et al., 2014) or by establishing new possibilities for high-throughput analysis, such as radiomics (Gillies et al., 2015). Over the past years, the improvements in automated brain tumor segmentation methods led to a steady increase in performance. This increase has two main reasons. First, the amount of annotated data has increased, leading to larger and more diverse datasets. Second, the available segmentation methods have evolved rapidly, especially with deep neural networks, which can leverage vast amounts of data. Although the results are reported to be close or on par with human performance (Meier et al., 2016; Bakas et al., 2018), there are still concerns about the clinical acceptability due to lower levels of robustness when compared to humans (Bakas et al., 2018). Possible reasons of this

lack of robustness comprise the large variability of the imaging properties (e.g., different vendors, magnetic field strength, artifacts), and the intrinsic heterogeneity of brain tumors itself.

One promising direction to alleviate the problem of robustness is using uncertainty estimates of automated segmentation results. In segmentation, where a class label is assigned to each voxel, the uncertainty typically reflects the confidence level of the predicted class label. In that sense, uncertainty estimates provide additional information on a method's prediction and might be employed in various ways, e.g., as visual feedback, to guide or automate corrections via segmentation error localization, or for segmentation failure detection at the patient level (i.e., systems outputting a single estimate reflecting the quality of the automated segmentation). Methods producing uncertainty estimates for neural networks exist for over 20 years (MacKay, 1992; Neal, 1995) and evolved steadily (Blundell et al., 2015; Hernández-Lobato and Adams, 2015) but have only recently been adapted for large and complex deep models, such as those employed for brain tumor segmentation. The most popular methods are: (a) Monte-Carlo (MC) dropout proposed by Gal and Ghahramani (2016), (b) aleatoric uncertainty estimation introduced by Kendall and Gal (2017), and (c) uncertainty from ensembles as shown by Lakshminarayanan et al. (2017). Their popularity is mainly due to their ability to be used with state-of-the-art segmentation methods, requiring only minor modifications to architecture and training.

The additional information provided through the uncertainty estimates might be employed to quantify the segmentation performance or as a post-processing step to correct automatic segmentations. Being able to reliably quantify the segmentation performance is crucial when using uncertainty estimates in clinical applications. Roy et al. (2019) and Wang et al. (2019) quantified the segmentation performance at structure level by using structure-wise uncertainty estimates as a proxy to predict the Dice coefficient of automated segmentation results. Similarly, Eaton-Rosen et al. (2018) obtained improved calibration accuracy and more reliable confidence intervals of brain tumor volume estimates from structure-wise uncertainty. The segmentation quality can also be assessed at subject level, which is of interest in clinical applications to flag possible failure cases for expert review. For brain tumor cavity segmentation (Jungo et al., 2018b) did so by aggregating voxel-wise uncertainty. In skin lesion segmentation (DeVries and Taylor, 2018) proposed to train a separate model predicting the segmentation's Dice coefficient based on the input image, the automated segmentation result, and the voxel-wise uncertainty estimates. Further, the uncertainty estimates can be used to correct automated segmentations. Nair et al. (2018) and Graham et al. (2019) showed improved results by using uncertainty estimates to exclude highly uncertain multiple sclerosis lesions and glands, respectively. Both works exclude structures based on uncertainty and thus use task-related knowledge (e.g., multi-lesion segmentation). Directly correcting voxel predictions based on uncertainty is not suggested since this requires to overrule the segmentation model that was optimized to perform the segmentation task. This is especially true when segmentation and uncertainty estimates are provided by the same model.

Although uncertainty estimation methods have been applied to different segmentation tasks, little is known on their usefulness and limitations, nor a common evaluation of their quality has been reported for medical image segmentation. Therefore, we analyzed the most commonly used uncertainty estimation methods in regards to benefits and challenges for brain tumor segmentation, which is one promising clinical application for computer-assisted medical image segmentation. We considered the methods' calibration, their segmentation error localization, and their segmentation failure detection ability (see **Figure 1** for an overview). This work builds on our previous work on the quality of uncertainties in medical image segmentation (Jungo and Reyes, 2019) and it is extended here in three aspects. First, based on our findings on observed deficiencies of voxel-wise uncertainty estimation approaches, we extend the work with experiments focusing on subject-level aggregation of uncertainty estimates. Second, to increase the clinical relevance of the analyses, we built and evaluated all methods for all three brain tumor labels (contrary to a simplified whole-tumor segmentation approach). Third, based on our previous work on the links between segmentation performance and quality of uncertainty estimates (Jungo et al., 2018a), we performed an experiment analyzing the effect of the training dataset size on the quality of uncertainty estimates.

2. MATERIALS AND METHODS

2.1. Data

We used the BraTS 2018 training dataset (Menze et al., 2015; Bakas et al., 2017a,b,c, 2018) consisting of 285 subjects with high- and low-grade brain tumors. Each subject comprises images of the four standard brain tumor magnetic resonance (MR) sequences: T1-weighted (T1), T1-weighted post-contrast (T1c), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR). Additionally, each subject holds a manual expert segmentation of three tumor sub-compartments: edema (ED), enhancing tumor (ET), and necrotic tissue combined with non-enhancing tumor (NCR/NET). In the official BraTS evaluation, these sub-compartments are combined into three hierarchical labels: whole tumor, tumor core, and enhancing tumor. Whole tumor (WT) is a combination of all tumor sub-compartments (i.e., ED, ET, NCR/NET), tumor core (TC) combines ET and NCR/NET, and enhancing tumor (ET) is defined by the ET sub-compartment. Aiming at yielding uncertainty estimates for these hierarchical tumor regions, we combined the tumor sub-compartment labels into the hierarchical labels before the training of the automated segmentation models.

The BraTS 2018 dataset comes pre-processed; the subjects and MR images are co-registered to the same anatomical template, resampled to unit voxel size ($1\times1\times1$ mm^3), and skull-stripped. We additionally normalized each MR image subject-wise to zero mean and unit variance. For all our experiments, we subdivided the BraTS training dataset into a split of 100 training, 25 validation, and 160 testing subjects, stratified by the tumor grade.

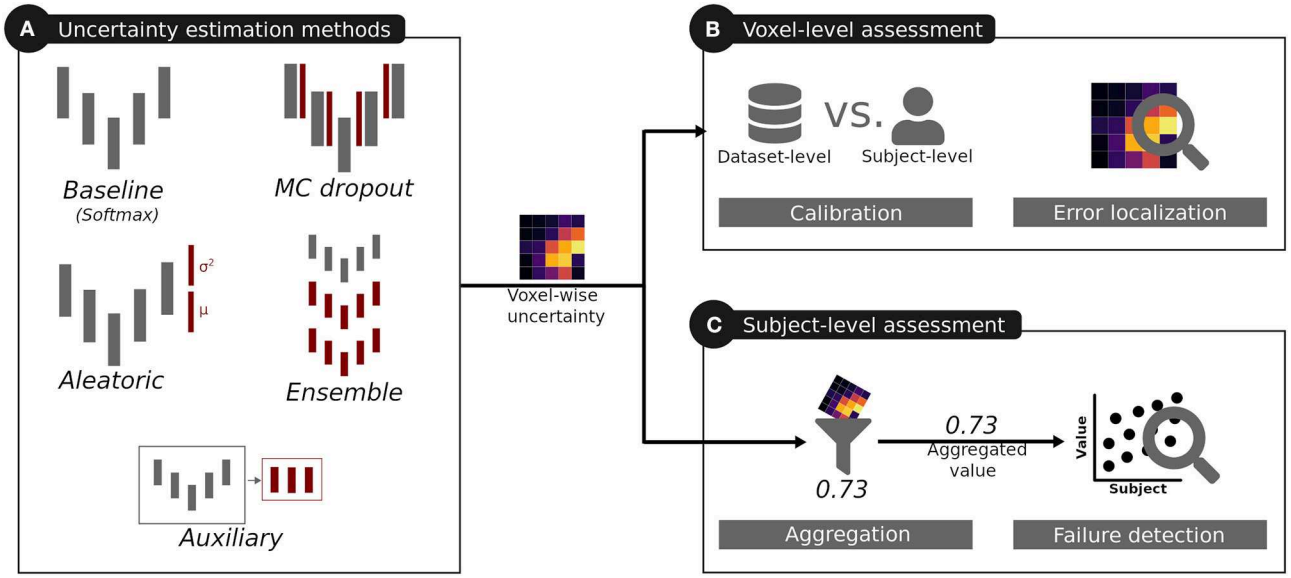


FIGURE 1 | Overview of the analysis performed for the uncertainties produced by different uncertainty estimation methods. The red color indicates additions introduced by these methods with respect to the baseline.

2.2. Experimental Setup

We used U-Net-like (Ronneberger et al., 2015) architectures to asses uncertainty estimation methods. The reason for using U-Net-like architectures is twofold. First, the widely used U-Net-like architectures are still state-of-the-art in brain tumor segmentation (Isensee et al., 2018; Myronenko, 2018) and, second, their simplicity minimizes architectural influences in the uncertainty estimates. Inspired by Nikolov et al. (2018), our U-Net processes anisotropic subvolumes of five consecutive axial slices to predict the corresponding center slices. As in Nikolov et al. (2018), we adopted a full-slice view which motivated us to use 2D+1D convolutions (i.e., 2D in-plane convolution followed by 1D out-plane convolution) instead of using 3D convolutions. By considering only the valid part of the convolution, each 1D convolution in the encoder part thereby reduces the off-plane size by two, leading to a fully 2D decoder. The architecture consists of four pooling/upsampling steps with two convolutions for each encoder and decoder level. Every convolution is followed by dropout ($p = 0.05$) (Srivastava et al., 2014), batch normalization (Ioffe and Szegedy, 2015), and ReLU activation (Glorot et al., 2011). The architecture has four input channels corresponding to the four MR images (i.e., T1, T1c, T2, FLAIR) and three sigmoid outputs, one for each of the three tumor regions (i.e., WT, TC, ET). We note that a single softmax output that includes all labels is prohibited by the hierarchy of the tumor regions. A detailed description of the network architecture can be found in the **Supplementary Section 1.1**. Adaptations of the presented architecture to the individual uncertainty estimation methods are described in section 2.3).

We used a common training scheme for all uncertainty estimation methods. This scheme consists of training the network for 50 epochs, from where we selected the best performing models based on the mean Dice coefficient across labels on the validation set. Furthermore, we used Adam optimizer (Kingma and Ba, 2015) (learning rate: 10^{-4}, β_1: 0.9, β_2: 0.999, ε: 10^{-8}) to optimize the cross-entropy loss in mini-batches of 24. Extensive fine-tuning of the individual methods might introduce large differences in segmentation performance. Therefore, in order to minimize the influence of the segmentation performance on the uncertainty estimates, we purposely omitted extensive fine-tuning of the individual methods. Likewise, we did not perform any data augmentation to reduce possible influences on the uncertainty estimates.

2.3. Uncertainty Estimation Methods

For our experiments, we considered five methods (**Figure 1A**) producing voxel-wise uncertainty estimates: one baseline, three common methods, and one auxiliary approach. The three common methods were selected due to their popularity in medical image segmentation, stemming from their simple integration into state-of-the-art segmentation methods.

2.3.1. Baseline: Softmax/Sigmoid Uncertainty

Although the softmax/sigmoid output is arguably a probability measure (Gal and Ghahramani, 2016), it is implicitly produced by classification networks. Therefore, we considered it as a reference comparison and named it *baseline*. We used the normalized entropy

$$\mathcal{H} = -\left[p_r log\left(p_r\right) + (1 - p_r) log\left(1 - p_r\right)\right] \frac{1}{log(2)} \in [0, 1] \quad (1)$$

as a measure of uncertainty, with p_r being the sigmoid output of the network (see section 2.2) for tumor region r.

2.3.2. MC Dropout

As shown by Gal and Ghahramani (2016), test-time dropout can be interpreted as an approximation of a Bayesian neural network. If applied during test time, dropout creates stochastic network samples that can be viewed as Monte-Carlo samples drawn from the posterior distribution of the network's weights. The foreground probability p_r of the tumor region r can be obtained by

$$p_r = \frac{1}{T}\sum_{t=1}^{T} p_{r,t}\,,$$

where T is the number of samples. We used the normalized entropy (Equation 1) as a measure of uncertainty.

We considered four different dropout strategies. The first strategy consists of applying MC dropout throughout all layers (see presented architecture in section 2.2), whose minimal dropout ($p = 0.05$) was intended as regularization. The second strategy is inspired by existing work in segmentation uncertainty (Kendall et al., 2015; Nair et al., 2018), where dropout is applied only at key positions. Accordingly, we modified the architecture (see section 2.2) and applied a prominent dropout ($p = 0.5$) at the center positions of the U-Net architecture only, i.e., before the pooling and after the upsampling operations (cf. illustration of MC dropout in **Figure 1A**). The third strategy is similar to the second but introduces the center dropout ($p = 0.5$) only at the two lowest pooling/upsampling steps. In a fourth strategy, we replaced the dropout of the initial architecture by concrete dropout (Gal et al., 2017). Concrete dropout learns the dropout probability as part of the optimization procedure and can, as the standard dropout, also be applied during test time to generate stochastic network samples. We refer to this four strategies as *baseline+MC*, *center+MC*, *center low+MC*, and *concrete+MC*. We considered the non-MC counterparts *center*, *center low*, and *concrete* as additional softmax/sigmoid uncertainties next to *baseline* (described in section 2.3.1).

2.3.3. Aleatoric Uncertainty

Aleatoric uncertainty is said to capture the noise inherent to an observation (Kendall and Gal, 2017) and is thus different from model uncertainty (e.g., MC dropout), which accounts for uncertainty in the model parameters. Kendall and Gal (2017) showed that aleatoric uncertainty in classification problems can be obtained by defining a network $f(x)$ for input x that generates two outputs

$$[\hat{x}, \sigma^2] = f(x)\,,$$

where $\hat{x}$ correspond to the logits, and σ^2 defines the variance of their Gaussian perturbation ($\mathcal{N}(\hat{x}, \sigma^2)$). The logits and the variance are simultaneously optimized by the aleatoric loss, which approximates the intractable objective with Monte-Carlo samples of the perturbed logits. We refer to this method as *aleatoric* and modified the architecture (see section 2.2) to output the variance σ_r^2 in addition to the logits $\hat{x}_r$ for every tumor region r (see **Supplementary Section 1.1** for a detailed architecture description). We used the $\hat{x}_r$ outputs for the segmentations and the σ_r^2 outputs as measures of uncertainty. To normalize the range of the variance maps across tumor regions, we normalized it to $[0, 1]$ over all subjects.

2.3.4. Ensembles

Ensembles of neural networks are typically used when performance is highly relevant, e.g., for the BraTS challenge (Kamnitsas et al., 2017), but they can also be used to quantify uncertainties (Lakshminarayanan et al., 2017). Our ensemble consists of $K = 10$ models that share the same architecture (see section 2.2) but differ in training to enforce variability. We trained each model k on alternating $K - 1$ folds of the training dataset (as in k-fold cross-validation, resulting in 90 instead of 100 training subjects. We obtained the foreground probability p_r for each tumor region r by the average

$$p_r = \frac{1}{K}\sum_{k=1}^{K} p_{k,r}$$

of all models. As an uncertainty measure we used the normalized entropy (Equation 1).

2.3.5. Auxiliary Networks

We use the term auxiliary network to describe additional networks that are trained successively to the primary network (i.e., segmentation network). Such networks have been used to assess segmentation performances by regressing subject-level performance metrics (DeVries and Taylor, 2018; Robinson et al., 2018). Inspired by this idea, we applied auxiliary networks for voxel-wise prediction of the segmentation errors (i.e., false positives, false negatives) of each tumor region separately. Since the auxiliary networks learn to detect segmentation errors, we can directly use their sigmoid segmentation error probabilities as a measure of segmentation uncertainty. Producing uncertainty estimates by a separate network is motivated by the presumption that a network might not be the best in assessing its own trustworthiness (Jiang et al., 2018).

We considered two types of auxiliary networks in our experiments. The first type, named *auxiliary feat.*, uses the features maps of the segmentation network (see section 2.2) as input and consists of three consecutive 1×1 convolutions. The second type, named *auxiliary segm.*, employs the three label maps (WT, TC, ET) produced by the segmentation network in combination with the four MR images as input. The difference between the two types of auxiliary networks is defined by the link to the segmentation network. The first is more closely linked through the feature maps, whereas the second is decoupled and only requires the resulting segmentations. We refer to the **Supplementary Section 1.2** for a detailed description of the *auxiliary feat.* and *auxiliary segm.* architectures.

2.4. Analyzing Voxel-Wise Uncertainty

We selected three techniques to analyze the quality of voxel-wise uncertainties produced by the different uncertainty estimation methods (**Figure 1B**) independently of their expressed uncertainty (e.g., model uncertainty, data uncertainty). The techniques aim at evaluating the model's confidence levels and the segmentation error localization abilities, which are required

for tasks relying on visual feedback, or guided/automated correction. Additionally, the Dice coefficient was used to monitor the segmentation performance.

2.4.1. Reliability Diagram

Reliability diagrams (DeGroot and Fienberg, 1983) assess the quality of a model's confidence. It is a visual measure of how close a model's calibration is to practically unachievable perfect calibration (Guo et al., 2017). Perfect calibration is obtained when a model's predictions $f(x)$ with confidence p are correct with a rate of p for any label y

$$P(y(x) = y|f(x) = p) = p\,,$$

where $y(x)$ are the model's label predictions. For instance, when a model is confident with 70%, it should be correct 70 out of 100 times (Guo et al., 2017). To create a reliability diagram, the continuous predictions $f(x)$ are discretized in M confidence bins c_m for $m \in \{1, \ldots, M\}$ and plotted against the accuracies a_m in these bins. Therefore, the identity line of the reliability diagram represents perfect calibration.

For segmentation tasks, the reliability diagrams are typically reported over an entire test set, jointly considering the confidences of all voxels across subjects. Although this offers a general idea of the model's overall calibration, it omits information about a single subject (i.e., patient). Achieving good calibration levels at subject-level is, however, required in a clinical setting if the voxel confidences should be used for visual feedback or guided corrections of automated segmentation results. Therefore, we report subject-level calibration along with dataset-level calibration.

Calibration builds on model confidence, which we used as a surrogate for uncertainty (as in Kendall and Gal, 2017). This consists in considering the tumor region probability p_r for the baseline, MC dropout and ensemble variants. Since aleatoric and auxiliary variants do not explicitly output probabilities p_r, we translated their uncertainty by $y(1 - 0.5q) + (1 - y)0.5q$ to confidence values, where $y \in \{0, 1\}$ is the segmentation label and $q \in [0, 1]$ is the normalized uncertainty.

2.4.2. Expected Calibration Error

The expected calibration error (ECE) (Naeini et al., 2015) distills the information of a reliability diagram into one scalar value. It is defined by the absolute calibration error between the confidence and accuracy bins, c_m and a_m, respectively, weighted by the number of samples n_m (in our case voxels) in the bin. More formally, with N and M being the total number of samples and the number of bins, the ECE is given by

$$ECE = \sum_{m}^{M} \frac{n_m}{N} |c_m - a_m|\,.$$

The ECE ranges from 0 to 1, where a lower value represents a better calibration. Through weighting by the bin size, the ECE is influenced by large confident and accurate extra-cranial regions typically found in brain tumor MR images. To reduce this effect, we only considered voxels within the skull-stripped brain to calculate the ECE. As for the reliability diagram, we are interested in the subject-level ECE and thus report the mean subject ECE instead of the dataset ECE (i.e., considering all voxels in the test set to calculate a single ECE). Complementary to the ECE, we also computed the average calibration error (Neumann et al., 2018). We refer to the **Supplementary Section 4** for the description and results.

2.4.3. Uncertainty-Error Overlap

In segmentation, not only calibration is of interest but also the model's ability to localize segmentation errors. Ideally, a model would be uncertain only where it makes mistakes. To assess this behavior, we introduce the uncertainty-error overlap (U-E). The U-E measures the overlap, through Dice coefficient, between the regions where the model is uncertain U about its prediction and the segmentation error E (i.e., union of false positives and false negatives), such that

$$U\text{-}E = \frac{2|U \cap E|}{|U| + |E|}\,,$$

where $|\cdot|$ represents the cardinality. The U-E ranges from 0 to 1 with 1 describing a perfect overlap. By considering voxels belonging to U and E only, the U-E is not influenced by the true negative uncertainty and thus typically independent of the image size or additional background voxels, as opposed to the ECE. However, calculating U-E requires to threshold U. We determined the threshold for each method independently, based on the maximal U-E performance on the validation set. The U-E performance was evaluated for thresholds from 0.05 to 0.95 in steps of 0.05. Complementary to the U-E, we also computed the area under the curve of the precision-recall curve. We refer to the **Supplementary Section 4** for the description and results.

2.4.4. Dice Coefficient

Although the Dice coefficient is not a measure for analyzing the quality of the uncertainty, we used it to monitor the segmentation performance of the different methods. It measures the overlap between two segmentation and ranges from 0 to 1, where 1 describes perfect overlap. Rather than determining the best method for segmentation, the Dice coefficient monitoring aims at detecting potential influences of the segmentation performance on the uncertainty estimates. Ideally, all methods would produce identical segmentations attributing any improvement in uncertainty measures directly to the corresponding method. In practice, however, this is unfeasible due to differences in the architectures and training. An improvement in the uncertainty measures could, therefore, also be due to an improved segmentation performance.

2.5. Analyzing Aggregated Uncertainty at the Subject Level

Besides analyzing the method's uncertainty estimates on a voxel level, we further analyzed their quality when aggregated on a subject level (i.e., one scalar value per subject; **Figure 1C**). The motivation of the subject-level analysis is twofold. First, the aggregation distills the uncertainty information such that the influence of irrelevant and erroneous voxel-wise information

is reduced. The aggregation, therefore, provides an assessment of the individual uncertainty estimations at a higher level that can forgive deficiencies (e.g., poor calibration) at the voxel level. Second, the aggregation presents a possible usage of the uncertainty estimations. It is an alternative to corrections at the voxel-level which are unfeasible for brain tumor segmentation where task-related knowledge (e.g., multiple lesions) is very sparse. In clinical applications subject-level information is important to flag possible failure cases for expert review. The vast amount of possible aggregations can further help in pointing to important characteristic of the voxel-wise uncertainty when used at the subject level. The quality of the aggregated subject-level information is defined by its relation to the segmentation performance; the better the aggregated uncertainty, the better it should be able to describe the segmentation performance. We aim at a good correlation between aggregated uncertainty and segmentation performance, which consequently enables accurate segmentation failure detection.

2.5.1. Aggregation Methods

The aggregated subject-level scalar is highly influenced by the chosen aggregation method. Hence, we studied three distinct aggregation methods.

Mean aggregation. Mean aggregation is one of the simplest aggregation methods, and it is motivated by the intuition that an overall higher voxel-wise uncertainty should be an indicator of poor segmentation performance. This requires the aggregated, but not necessarily the voxel-wise, uncertainty to be calibrated. In practice, we used the negative mean uncertainty to obtain direct relation to the segmentation performance.

Prior knowledge-based aggregation. We know that uncertainty is inherently present at the segmentation boundary. Although this boundary uncertainty might be well-calibrated it is mainly proportional to the size of the segmentation and consequently introduces a bias toward the tumor size to the aggregation. Similarly, one might expect more severe issues when the large amount of uncertainties are present far from the segmentation boundary. If only boundary uncertainty is present we would expect less deviation from a reference segmentation. We used this knowledge to create three different aggregation weightings which deemphasize uncertainty at boundaries. The first weighting consists of masking out voxels at the segmentation boundary. In our experiments we masked three voxels within the boundary (i.e., one-pixel distance inside and outside, and at the boundary). The second weighting considers the distance to the boundary, penalizing uncertainties close to the boundary, and up-weighting uncertainties distant from the segmentation boundary. The third weighting normalizes the boundary uncertainty by dividing through the segmentation volume.

To aggregate the differently weighted voxel information to a scalar value per subject, we used three simple operations: mean, sum, and logsum (as used by Nair et al., 2018). We considered nine combinations between prior knowledge-based weightings and these three simple operations. The nine combinations were then used to train a random forest regressor that predicts the Dice coefficient of the segmentation. We used such a prediction model instead of evaluating the correlation with the segmentation performance because we aim at obtaining a good predictor rather than solely finding the most important combination. We refer to the **Supplementary Sections 2.1, 2.3** for details regarding the nine combinations and training details of the random forest regressor.

Aggregation with automatically-extracted features. Instead of manually defining additional aggregation methods, we employed the `PyRadiomics`[1] (Van Griethuysen et al., 2017, version 2.2.0) package to extract subject-level features from the voxel-wise uncertainty estimates automatically. Although typically used in the context of radiomics, the package is not limited to this application but is rather a general tool to extract shape, first-order, and other gray-level features. The benefit of using automated feature extraction is two-fold: (a) it allows us to compare to the aggregation with prior knowledge and (b) potentially points to new predictive features of the uncertainty. We extracted 102 features from the thresholded voxel-wise uncertainty estimates. The threshold was determined for each uncertainty method by the maximal U-E performance on the validation set (identical to section 2.4.3). The features were used to train a random forest regressor that predicts the Dice coefficient of the segmentations. We refer to the **Supplementary Sections 2.2, 2.3** for features and training details.

2.5.2. Subject-Level Metrics

We assessed the three aggregation methods for each uncertainty estimation method based on their ability to predict the Dice performance of the automated segmentations. To do so, we evaluated the estimates of the aggregation methods by three metrics.

Spearman's rank correlation. We used Spearman's rank correlation coefficient to asses the correlation between the estimated and the actual Dice coefficients. Spearman's rank correlation was chosen since not all estimates lead to a linear relationship (i.e., mean aggregation). The metric ranges from -1 to 1, where the extremes describe a perfect monotone relation (positive if 1, negative if -1) between estimated and actual Dice coefficients.

AUC-ROC. We evaluated the segmentation failure detection abilities of the uncertainty and aggregation methods by the area under the curve of the receiver operating characteristic (AUC-ROC). To do so, we translated the regression problem (i.e., building a predictor for the Dice coefficient) to a binary classification problem. We classified the segmentations in successful and failed according to the average inter-rater Dice coefficient for every tumor region. As the inter-rater performances are not provided for the BraTS 2018 dataset, we considered the inter-rater performances reported in Menze et al. (2015) for the BraTS 2013 dataset[2]. The AUC-ROC was

[1] https://pyradiomics.readthedocs.io

[2] Note that the inter-rater performance between the BraTS 2013 and the BraTS 2018 dataset might differ since the existing annotations were revised by expert board-certified neuroradiologists. Also, the non-enhancing and the necrosis label were fused.

computed by the scores of the regression output and ranges from 0 to 1, where 1 describes a perfect separator between the classes, 0.5 corresponds to random guesses, and 0 is the reciprocal of a prefect separator (i.e., consistently predicting wrong class).

Youden's accuracy. For improved comparability and understanding, we evaluated the accuracy (range $[0, 1]$) along with the AUC-ROC. We used the maximal Youden's index (Youden, 1950) to determine the accuracy from the ROC curve. This index is defined for each point on the ROC curve as

$$J = sensitivity - (1 - specificity)$$

and corresponds to the vertical distance to the chance line (i.e., $sensitivity = 1 - specificity$). Its maximum defines an optimal point on the ROC curve.

3. RESULTS

3.1. Dataset-Level vs. Subject-Level Calibration

Figure 2 illustrates the difference between dataset-level (i.e., all voxels in the test set) and subject-level (i.e., all voxels of a subject) calibration with reliability diagrams. While the calibration at the dataset level is good for all tumor regions, miscalibrations in the form of overconfidence and underconfidence are present at the subject level. We find an under-/overconfidence in 39%/25%, 30%/32%, and 21%/41% of the test subjects for the three tumor regions WT, TC, and ET. Consequently, less than 40% (36%, 38%, and 38%) of the subjects are well-calibrated. The percentages indicate that the amount of miscalibration is similar for all tumor regions, but ET exhibits more overconfidence (and less underconfidence) than the other regions (column *underconfident subject* in **Figure 2** is exemplary). Also, we observe small differences among the uncertainty methods; they mostly agree, except for the aleatoric uncertainty, which disagrees at the dataset level.

3.2. Voxel-Wise Uncertainty

We evaluated the voxel-wise uncertainties on average subject-level ECE, uncertainty-error overlap (U-E), and Dice coefficient. The results are listed in **Table 1** and reveal that no uncertainty estimation method considerably outperforms others. Most methods perform in a similar range with a small advantage for the *ensemble* method. Only the *aleatoric* method is distinctly performing worse in term of the uncertainty metrics ECE and U-E, while the competitive Dice coefficients indicate no segmentation related issues. We found that the MC dropout variants typically marginally outperform the non-MC variants (i.e., dropout only applied during training), but occasionally lead to considerable gains in ECE. The results also show that finding the optimal dropout strategy, i.e., the amount and position of dropout, is not evident. On one hand, the method containing moderate dropout (*center low/+MC*) outperforms the methods with minimal (*baseline/+MC*) and maximal (*center/+MC*) dropout on all metrics. On the other hand, the benefit of using MC dropout is larger for the minimal and maximal dropout strategies. Concrete dropout, which learns an optimal dropout rate, yielded comparable but not superior results than (*center low/+MC*). Furthermore, the results show that the auxiliary methods achieved uncertainty performances on par with the *baseline* model, on whose segmentation errors they are trained. Benefits in comparison to *baseline* are mainly found for *auxiliary feat.* and in terms of U-E.

Overall, the results are similar for all tumor regions. Differences among the tumor regions are mainly found in the ECE metric, which is considerably lower for ET than WT. This effect can be explained since the ET includes substantially fewer voxels predicted as uncertain (since fewer foreground voxels) and, in turn, the ET tumor class includes more certain background voxels, leading to an improved ECE. Furthermore, the results indicate a link between segmentation performance and ECE, where better-performing methods often relate to an improved ECE. Methods outputting the uncertainty estimates separately from the segmentation (i.e., *auxiliary segm.*, *auxiliary feat.*, and *aleatoric*) are excluded from this observation.

Figure 3 shows the uncertainty estimates for the WT label (see **Figures S3, S4** for visual examples of TC and ET) produced by the selected methods on underconfident, overconfident, and well-calibrated subjects (same subjects as in **Figure 2**). The examples visually confirm the similar segmentation performances of the different methods. Further, the uncertainty estimates clearly show a pattern between amount of uncertainty and miscalibration. The underconfident subject exhibits considerably more overall uncertainty than the overconfident subject and perceivably more than well-calibrated subject. We also observe that the amount of uncertainty varies among the methods. For instance, the *center/+MC* methods consistently exhibit more uncertainty than the auxiliary methods. The regions exhibiting uncertainty are, however, similar for all methods, except the *aleatoric* method which visually confirms its poor calibrations.

In an additional experiment, we analyzed the dependency of the training dataset size on the quality of uncertainty estimates. The method we used for these experiments is *baseline+MC* as it is mostly represents the performance level of the studied methods. The results in **Figure 4** show that quality in terms of ECE is low (i.e., high ECE) with few training data and increases afterwards. This demonstrates that the higher uncertainty introduced through small datasets is worse in terms of quality.

3.3. Subject-Level Aggregated Uncertainty

Figure 5 shows the AUC-ROC results of the uncertainty estimation methods for the three aggregation methods (see **Figure S5** and **Table S5** for the corresponding ROC curves and the AUC-ROC values, respectively). The results demonstrate that mean aggregation has a limited ability to detect segmentation failures and is comparable with guessing (i.e., AUC-ROC of 0.5). Results for the ET tumor label are even below 0.5, revealing a direct relation between uncertainty and segmentation performance instead of the expected inverse relation. An improvement over mean aggregation is achieved by aggregating with prior knowledge and automatically extracted features. The aggregation with automatically extracted features obtained the

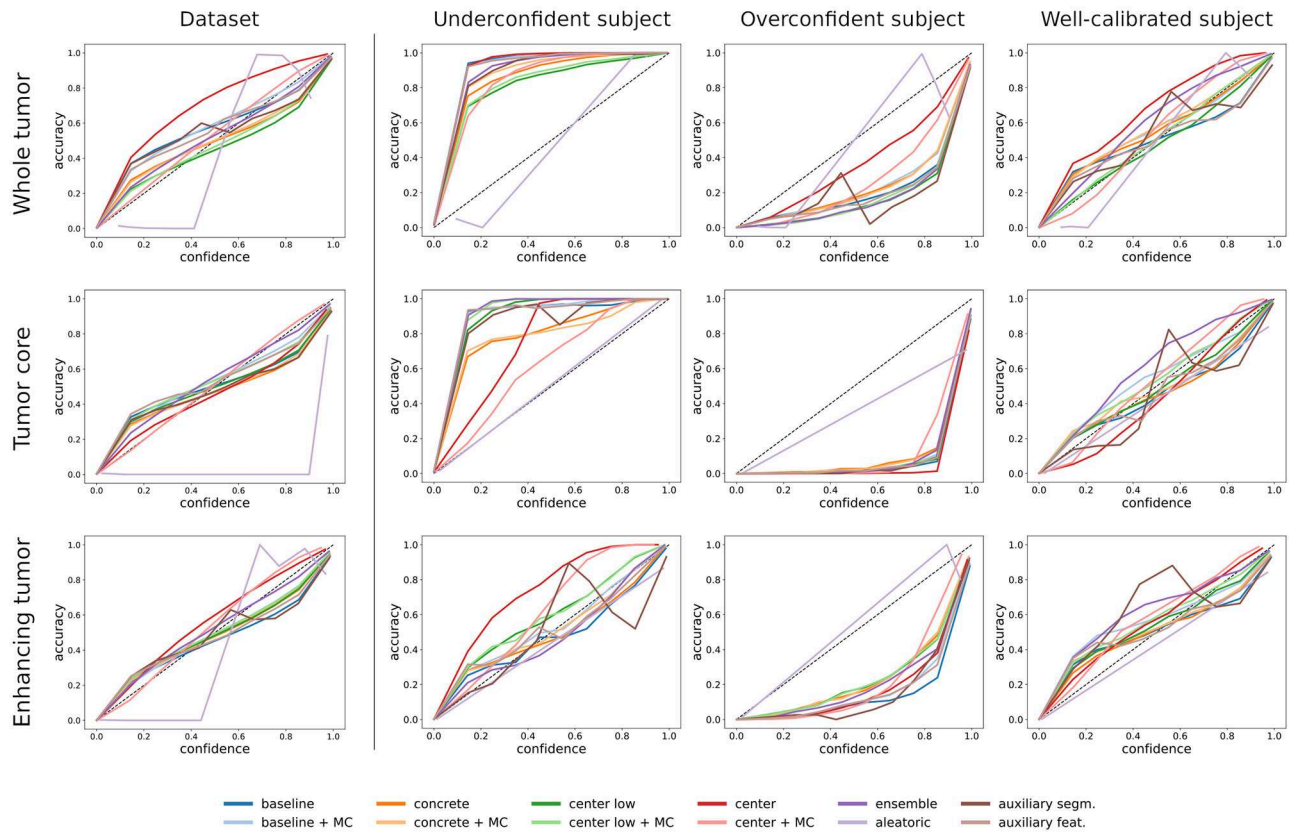


FIGURE 2 | Comparison between dataset-level and subject-level calibration (shown in reliability diagrams) for the selected uncertainty estimation methods. The first column shows the dataset-level calibration, which considers all voxels in the dataset. The second to fourth columns show subject-level calibrations, which consider voxels of a single subject. The exemplary subjects indicate underconfident, overconfident, and well-calibrated methods. The rows indicate the three tumor regions.

TABLE 1 | Performances of the different uncertainty estimation methods in terms of expected calibration error (ECE), uncertainty-error overlap (U-E), and Dice coefficient.

	WT			TC			ET		
	ECE%	**U-E**	**Dice**	**ECE%**	**U-E**	**Dice**	**ECE%**	**U-E**	**Dice**
Baseline	1.059	0.427	0.869	0.853	0.41	0.767	0.309	0.401	0.692
Concrete	0.984	0.429	0.875	0.802	0.419	0.775	0.278	0.407	0.686
Center low	0.942	0.434	0.88	0.83	0.409	0.775	0.28	0.403	0.686
Center	1.606	0.425	0.817	1.086	0.41	0.695	0.381	0.395	0.642
Baseline + MC	1.016	0.433	0.869	0.805	0.41	0.765	0.284	0.403	0.693
Concrete + MC	0.952	0.431	0.877	0.785	**0.422**	**0.778**	**0.27**	0.409	0.689
Center low + MC	0.922	0.435	**0.881**	0.83	0.41	0.769	0.275	0.409	0.69
Center + MC	1.014	0.432	0.874	1.06	0.409	0.716	0.462	0.4	0.651
Ensemble	**0.893**	**0.436**	0.88	**0.749**	0.402	**0.778**	0.275	0.411	**0.701**
Aleatoric	12.187	0.001	0.874	2.407	0	0.757	1.284	0.007	0.673
Auxiliary segm.	1.058	0.428	0.869	0.887	0.397	0.767	0.323	0.39	0.692
Auxiliary feat.	1.057	0.433	0.869	0.852	0.403	0.767	0.318	**0.423**	0.692

All metrics range from 0 to 1, but the ECE is reported in % for better comparisons. Lower ECEs are better as well as higher U-Es and Dice coefficients.We note that the Dice coefficient is not a measure for analyzing the quality of the uncertainty and is reported to monitor the segmentation performance of the different methods. Mean values are presented, and standard deviations are omitted due to marginal differences. Bold values indicate best performances. Horizontal separations group types of uncertainty methods and WT, TC, and ET indicate the tumor regions whole tumor, tumor core, and enhancing tumor.

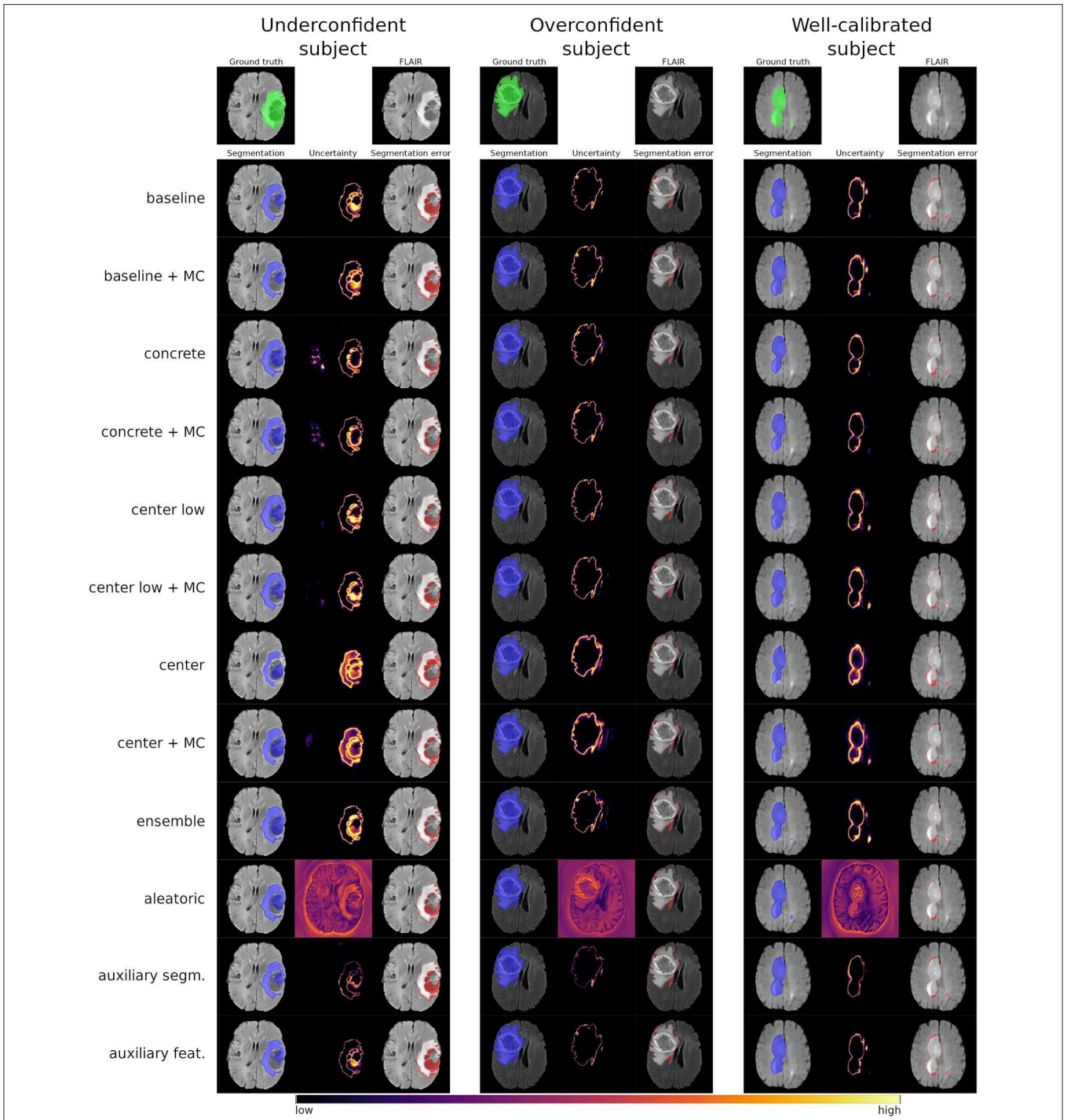


FIGURE 3 | Visual examples of the whole tumor uncertainty produced by the different uncertainty estimation methods. The columns correspond to underconfident, overconfident, and well-calibrated subjects (same as in **Figure 2**).

overall best AUC-ROC values. Although it is not possible to determine the best uncertainty method visually, the *aletoric* method shows apparent weaknesses. We also built a combined model with the automatically extracted features and the generated prior knowledge features, but it did not lead to consistent improvements in terms of AUC-ROC.

We assessed feature importance by accumulating their ranks over the individual regression models (i.e., one for each uncertainty estimation method). We found the *distance weighted masked mean* feature to be the most important feature for the prior knowledge-based aggregation. For the aggregation with automatically extracted features, the *shape sphericity*, which is

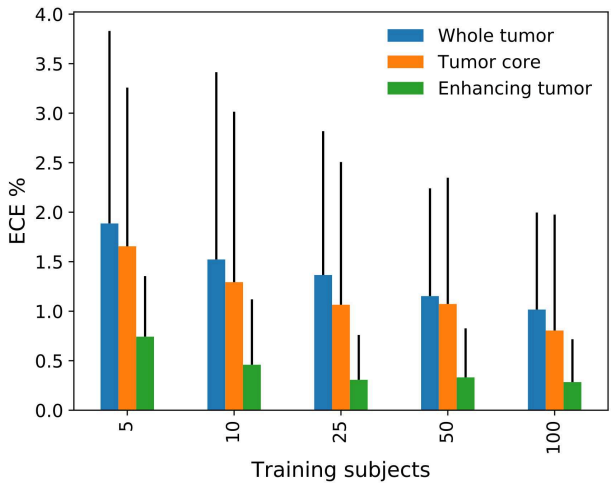


FIGURE 4 | Effect of the training dataset size on the expected calibration error (ECE).

a measure of roundness relative to a sphere, and *run length non-uniformity*, which is a measure of similarity among the different gray level run lengths, were dominantly the two most important features.

Since it represents the best-performing aggregation method, we evaluated the aggregation with automatically extracted features in terms of Spearman's rank correlation and Youden's accuracy in addition to AUC-ROC. The corresponding results are shown in **Figure 6** with numerical details in **Table S5**. The results reconfirm the similarities found among the different uncertainty estimation methods, with the *aleatoric* method yielding the lowest performance, and producing negative outliers in all three metrics, although less prominent for the Youden's accuracy. Additionally, we observed that the predictions based on the TC tumor label uncertainty achieved the highest values for all three metrics, whereas ET tumor label uncertainty is typically the worst-performing.

4. DISCUSSION

Uncertainty estimation methods have been used for different medical image segmentation tasks, but little is known on their quality and limitations. Therefore, we analyzed the quality of common uncertainty estimation methods on the clinically relevant problem of automated brain tumor segmentation. The methods were evaluated on their calibration, segmentation error localization, and segmentation failure detection abilities. First, our results show that overall good calibration is only achieved at the dataset level. Second, segmentation error localization relying on voxel-wise uncertainty is difficult and unreliable. However, we found that segmentation failure detection on subject level is possible by aggregating voxel-wise uncertainty estimates.

We found a good calibration of voxel-wise predictions at the dataset level but observed notable miscalibrations when assessed at the subject level. As such, the good dataset-level calibration can be explained by the subject-level miscalibrations (under- and overconfidence), which average out when combined. The subject-level miscalibrations are influenced by the dependence of neighboring voxels in the uncertainty estimates, resulting from the fully-convolutional architectures and an inter-voxel dependence in the MR images itself. Although this dependence is beneficial for the segmentation task, it also produces similar uncertainties within a neighborhood and therefore biases the calibration. Consequently, poor segmentations are expected to introduce a larger bias. The observed miscalibrations further indicate that the uncertainty estimates may contain non-negligible errors. As a consequence, using voxel-wise uncertainty for user feedback or guided corrections is questionable and might lead to undesired outcomes in automated corrections. Therefore, our findings point on the importance of developing methods able to calibrate uncertainties for each subject individually.

The results of the voxel-wise uncertainty evaluation reveal that all methods (including the softmax/sigmoid baseline) performed similarly, except for the aleatoric uncertainty, which performed worst. Among the similar performing methods, the ensemble achieved the overall best results. It achieved improved uncertainty metrics along with its expected benefits in segmentation performance. MC dropout, which can be viewed as a poor man's ensemble equivalent, showed benefits similar to ensemble when compared to using standard dropout (i.e., during training only). However, finding the optimal dropout strategy that maximizes segmentation performance and uncertainty estimates remains difficult, also because concrete dropout showed not to be optimal. Therefore, as a rule of thumb, we suggest using ensembles when resources allow it. Otherwise, we suggest applying MC dropout with a focus on regularization benefits.

The results further indicate a possible relation between the quality of the uncertainty and segmentation performance. For instance, the ensemble and the MC dropout methods revealed benefits for the uncertainty along with improved segmentation performance. It is impossible to determine whether these methods are effectively producing qualitatively better uncertainties or the increased quality results from the improved segmentation. To assess the uncertainty separately, methods that produce decoupled uncertainty estimates would be required, as advocated by Jiang et al. (2018). Our auxiliary networks are examples of such decoupled solutions. They showed promising results but without achieving substantial benefits. Further work in this direction is needed to determine its full potential. The experiment with limited training data confirmed the observation of a link between segmentation performance and quality of the uncertainties by showing improved quality with increasing dataset size. This observation is troublesome because large datasets are rare and qualitatively good uncertainties would be especially desirable for underperforming models due to little training data.

Aggregating the voxel-wise uncertainty can distill valuable information for segmentation failure detection. The best-performing aggregation method tested was the aggregation with automatically extracted features. It achieved a good correlation with the Dice coefficient and enabled an accurate separation

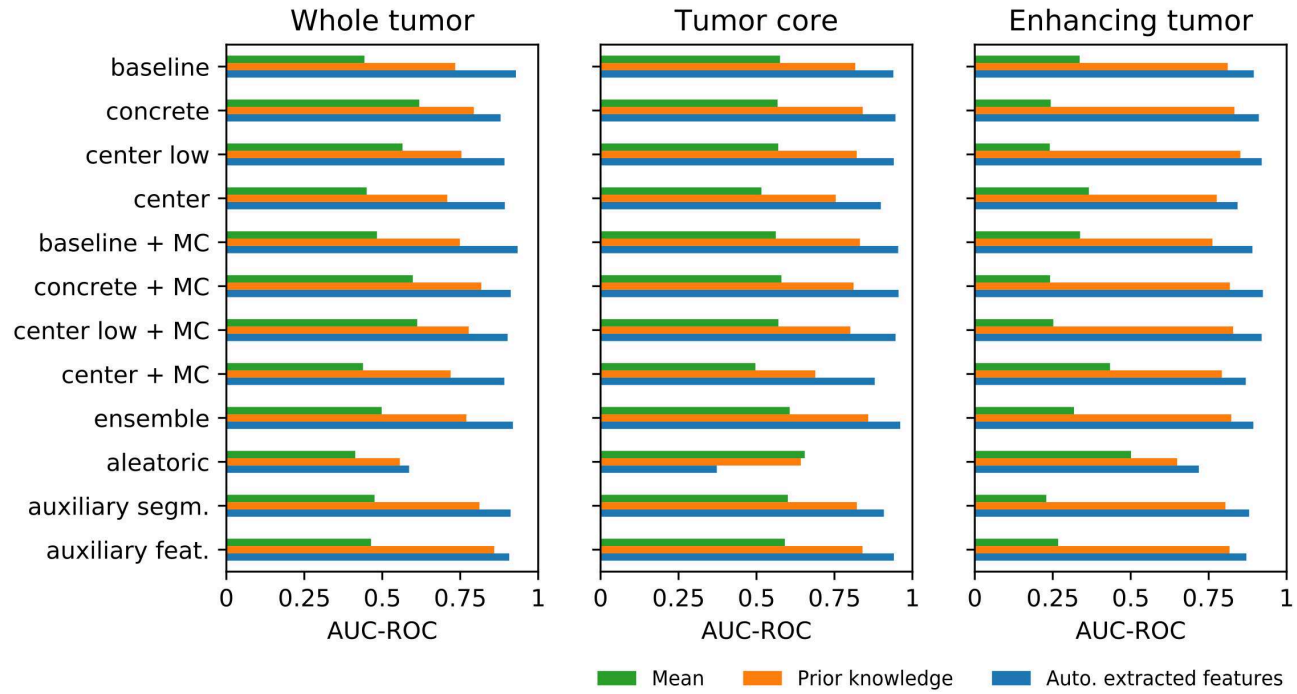


FIGURE 5 | Differences among the three aggregation methods for each uncertainty estimation approach in terms of area under the curve of the receiver operating characteristic (AUC-ROC) for segmentation failure detection.

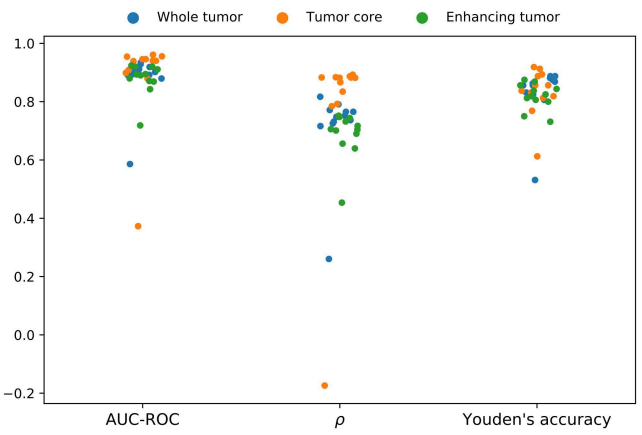


FIGURE 6 | Segmentation failure detection performance of the aggregation uncertainty by automatically extracted features in terms area under the curve of the receiver operating characteristic (AUC-ROC) and Youden's accuracy as well as correlation with the segmentation performance in terms of Spearman's rank correlation (ρ). Each point per color represents an uncertainty estimation method. The negative outliers in each metric and for each tumor region correspond to the *aleatoric* method.

between successful and failed segmentations results. For the mean aggregation, we obtained notably worse results, indicating a poor relation between mean uncertainty and segmentation performance. However, we could greatly improve this relation by simply weighting the uncertainties according to some prior knowledge. A subsequent feature importance analysis revealed that, particularly, the distance to the segmentation boundary matters as prior knowledge. For the aggregation with automatically extracted features, two important features in the voxel-wise uncertainty estimates were revealed: *shape sphericity* and *run length non-uniformity*. The importance of the sphericity can, to some extent, be explained by its definition, which consists of a ratio between mesh volume and surface area. This definition results in low sphericities for large-area-low-volume structures such as a narrow uncertainty rim that we would expect for a successful segmentation. Considering volume and area at the same time might be key to cope with the highly variable brain tumor volumes and areas. Similarly, the narrow uncertainty rim of successful segmentations is expected to contain a lot of similar uncertainty levels and thus resulting in a lower run length non-uniformity than of failed segmentations.

Overall, the analyzed uncertainty estimation methods only limitedly provide the desired additional and useful information. Our results question whether a remedy of the challenges with voxel-wise uncertainties is even feasible. Additional processing is required to take advantage of the voxel-wise estimates. We presented such an additional processing by aggregating the voxel-wise uncertainties into one value per subject and achieved promising results for segmentation failure detection. The promising aggregation results point in the direction of an intermediate approach, operating in-between voxel and subject level. We believe this is important in clinical applications where uncertainty estimation methods would directly operate at the levels of lesion, region, or image slice e.g., for automated segmentation correction.

Our evaluation has several limitations worth mentioning. First, due to its popularity we used a U-Net-like architecture with a shared learning scheme for all our experiments. Our findings may differ for other setups, especially when altering the output confidences of a network, such as Dice coefficient loss as shown by Sander et al. (2019). Second, the metrics used to analyze the quality are comparing with the ideal case. Although good metrics signify high quality, the opposite (i.e., bad metrics mean low quality) might not be true, since the quality is not solely defined by the employed metrics. Moreover, low metric results, as for the U-E, do not directly mean that the uncertainty information is useless but might require additional steps to create benefit. Third, we used a selection of commonly used uncertainty estimation methods. Hence, we cannot claim that these findings apply to other, recently proposed techniques (e.g., Baumgartner et al., 2019; Jena and Awate, 2019; Wang et al., 2019). Also, we analyzed the different uncertainty estimation methods independently of their expressed uncertainty (e.g., model uncertainty, data uncertainty). While this provides information on the quality across types of uncertainty, an independent analysis by type of uncertainty might bring additional insights for the development of new uncertainty estimation methods.

In conclusion, we analyzed common uncertainty estimation methods and found that the quality of their voxel-wise uncertainty is limited in terms of subject-level calibration and segmentation error localization. We further showed that aggregating the voxel-wise uncertainties to the subject level enables accurate segmentation failure detection, which after all confirms the usefulness of the uncertainty estimates. We suggest a careful usage of voxel-wise uncertainty measures and highlight the importance of developing solutions that address the subject-level requirements on calibration and segmentation error localization.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the CBICA Image Processing Portal https://ipp.cbica.upenn.edu/.

AUTHOR CONTRIBUTIONS

AJ, FB, and MR contributed conception and design of the study. AJ and FB contributed implementation of the method. AJ conducted the experiments. AJ and MR wrote the manuscript. All authors contributed to manuscript revision, proofreading, and approved the submitted version.

FUNDING

This work was supported by the Swiss National Science Foundation (SNF) by grant number 169607 and by the Swiss Personalized Health Network (SPHN) initiative.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00282/full#supplementary-material

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. *Cancer Imaging Arch.* doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *Cancer Imaging Arch.* doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv: 1811.02629*.

Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., et al. (2019). "Phiseg: capturing uncertainty in medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. (Cham: Springer), 119–127. doi: 10.1007/978-3-030-32245-8_14

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv: 1505.05424*.

DeGroot, M. H., and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *J. R. Stat. Soc. Ser. D* 32, 12–22. doi: 10.2307/2987588

DeVries, T., and Taylor, G. W. (2018). Leveraging uncertainty estimates for predicting segmentation quality. *arXiv: 1807.00502*.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., and Cardoso, M. J. (2018). "Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 691–699. doi: 10.1007/978-3-030-00928-1_78

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning, Vol. 48* (New York, NY: PMLR), 1050–1059.

Gal, Y., Hron, J., and Kendall, A. (2017). "Concrete dropout," in *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc.), 3581–3590.

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2015). Radiomics: images are more than pictures, they are data. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Glorot, X., Bordes, A., and Bengio, Y. (2011). "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Vol. 15* (Fort Lauderdale, FL: PMLR), 315–323.

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.-A., Snead, D., et al. (2019). Mild-net: minimal information loss dilated network for gland instance segmentation in colon histology images. *Med. Image Anal.* 52, 199–211. doi: 10.1016/j.media.2018.12.001

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 (PMLR), 1321–1330.

Hernández-Lobato, J. M., and Adams, R. (2015). "Probabilistic backpropagation for scalable learning of bayesian neural networks," in *Proceedings of the 32nd International Conference on Machine Learning, Vol. 37* (Lille: PMLR), 1861–1869.

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, Vol. 37* (Lille: PMLR), 448–456.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 234–244. doi: 10.1007/978-3-030-11726-9_21

Jena, R., and Awate, S. P. (2019). "A bayesian neural net to segment images with uncertainty estimates and good calibration," in *International Conference on Information Processing in Medical Imaging* (Cham: Springer), 3–15. doi: 10.1007/978-3-030-20351-1_1

Jiang, H., Kim, B., Guan, M., and Gupta, M. (2018). "To trust or not to trust a classifier," in *Advances in Neural Information Processing Systems 31* (Curran Associates, Inc.), 5541–5552.

Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., et al. (2018a). "On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 682–690. doi: 10.1007/978-3-030-00928-1_77

Jungo, A., Meier, R., Ermis, E., Herrmann, E., and Reyes, M. (2018b). "Uncertainty-driven sanity check: application to postoperative brain tumor cavity segmentation," in *International Conference on Medical Imaging with Deep Learning*.

Jungo, A., and Reyes, M. (2019). "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 48–56. doi: 10.1007/978-3-030-32245-8_6

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2017). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 450–462. doi: 10.1007/978-3-319-75238-9_38

Kendall, A., Badrinarayanan, V., and Cipolla, R. (2015). Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv: 1511.02680*.

Kendall, A., and Gal, Y. (2017). "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc.), 5574–5584.

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *International Conference for Learning Representations*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc.), 6402–6413.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural Comput.* 4, 448–472. doi: 10.1162/neco.1992.4.3.448

Meier, R., Knecht, U., Loosli, T., Bauer, S., Slotboom, J., Wiest, R., et al. (2016). Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Sci. Rep.* 6:23376. doi: 10.1038/srep23376

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Myronenko, A. (2018). "3d MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 311–320. doi: 10.1007/978-3-030-11726-9_28

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2018). "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 655–663. doi: 10.1007/978-3-030-00928-1_74

Neal, R. M. (1995). *Bayesian learning for neural networks*. (Ph.D. thesis). University of Toronto, Toronto.

Neumann, L., Zisserman, A., and Vedaldi, A. (2018). "Relaxed softmax: efficient confidence auto-calibration for safe pedestrian detection," in *NIPS Workshop on Machine Learning for Intelligent Transportation Systems*.

Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., et al. (2018). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv: 1809.04430*.

Reuter, M., Gerstner, E. R., Rapalino, O., Batchelor, T. T., Rosen, B., and Fischl, B. (2014). Impact of MRI head placement on glioma response assessment. *J. Neuro-Oncol.* 118, 123–129. doi: 10.1007/s11060-014-1403-8

Robinson, R., Oktay, O., Bai, W., Valindria, V. V., Sanghvi, M. M., Aung, N., et al. (2018). "Real-time prediction of segmentation quality," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 578–585. doi: 10.1007/978-3-030-00937-3_66

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., and Initiative, A. D. N. (2019). Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage* 195, 11–22. doi: 10.1016/j.neuroimage.2019.03.042

Sander, J., de Vos, B. D., Wolterink, J. M., and Išgum, I. (2019). "Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI," in *Medical Imaging 2019: Image Processing*, Vol. 10949 (International Society for Optics and Photonics; SPIE), 324–330. doi: 10.1117/12.2511699

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338, 34–45. doi: 10.1016/j.neucom.2019.01.103

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* 3, 32–35.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 16 April 2020
doi: 10.3389/fncom.2020.00032

Measuring Efficiency of Semi-automated Brain Tumor Segmentation by Simulating User Interaction

David Gering, Aikaterini Kotrotsou, Brett Young-Moxon, Neal Miller, Aaron Avery, Lisa Kohli, Haley Knapp, Jeffrey Hoffman, Roger Chylla, Linda Peitzman and Thomas R. Mackie*

HealthMyne Inc., Madison, WI, United States

Traditionally, radiologists have crudely quantified tumor extent by measuring the longest and shortest dimension by dragging a cursor between opposite boundary points across a single image rather than full segmentation of the volumetric extent. For algorithmic-based volumetric segmentation, the degree of radiologist experiential involvement varies from confirming a fully automated segmentation, to making a single drag on an image to initiate semi-automated segmentation, to making multiple drags and clicks on multiple images during interactive segmentation. An experiment was designed to test an algorithm that allows various levels of interaction. Given the ground-truth of the BraTS training data, which delimits the brain tumors of 285 patients on multi-spectral MR, a computer simulation mimicked the process that a radiologist would follow to perform segmentation with real-time interaction. Clicks and drags were placed only where needed in response to the deviation between real-time segmentation results and assumed radiologist's goal, as provided by the ground-truth. Results of accuracy for various levels of interaction are presented along with estimated elapsed time, in order to measure efficiency. Average total elapsed time, including loading the study through confirming 3D contours, was 46 s.

Keywords: brain MRI, tumor, segmentation, glioma, deep learning, efficiency

OPEN ACCESS

Edited by:
Spyridon Bakas, University of Pennsylvania, United States

Reviewed by:
Guotai Wang, University of Electronic Science and Technology of China, China
Tuo Zhang, Northwestern Polytechnical University, China

****Correspondence:***
David Gering
david.gering@healthmyne.com

Received: *26 November 2019*
Accepted: *24 March 2020*
Published: *16 April 2020*

Citation:
Gering D, Kotrotsou A, Young-Moxon B, Miller N, Avery A, Kohli L, Knapp H, Hoffman J, Chylla R, Peitzman L and Mackie TR (2020) Measuring Efficiency of Semi-automated Brain Tumor Segmentation by Simulating User Interaction. Front. Comput. Neurosci. 14:32. doi: 10.3389/fncom.2020.00032

INTRODUCTION

Malignant brain tumors often have unfavorable prognoses such as time to progression and overall survival, and also have direct impact on motor and/or cognitive function and poor quality of life (Omuro and DeAngelis, 2013). In recent decades, imaging has played a key role throughout the entire treatment paradigm of cancer patients ranging from diagnosis and presurgical planning to treatment response assessment. Additionally, multimodal MRI protocols allow for non-invasive interrogation of tumor heterogeneity and identification of phenotypic sub-regions i.e., peritumoral edema/invasion, enhancing active tumor core and necrotic regions which reflect tumor biological properties including tumor cellularity, vascularity, and blood-brain barrier integrity.

However, despite the exponential enhancement in imaging sequences, hardware and software, we have barely begun to tap the potential of non-invasive imaging to characterize the phenotype of tumors. To date, radiologic assessments are qualitative including tumor detection and image-based tumor staging or semi-quantitative using freehand uni-dimensional and

bi-dimensional measurements of the tumor. In fact, all current imaging assessment criteria [such as Response Evaluation Criteria in Solid Tumors (RECIST), Response Assessment in Neuro-Oncology (RANO), immune related RECIST (irRECIST), and immune related response criteria (irRC)] used to evaluate tumor response in the clinical setting or in clinical trials rely on these freehand measurements to evaluate tumor size (Sorensen et al., 2008; Eisenhauer et al., 2009; Wolchok et al., 2009; Wen et al., 2010).

Accurate assessment of tumor volume is important for clinical management and particularly for monitoring treatment response and development of new therapies and trials. Despite the well-known advantages of whole tumor volumetric assessment, as recognized by the RANO Working Group, currently it is only performed for research purposes as manual outlining can be time-consuming, and it is susceptible to inherent intra-observer and inter-observer variability (Wen et al., 2010). Research efforts have focused on the development of computer-aided techniques for tumor segmentation. Computer-aided tumor segmentation techniques can be grouped in two major categories based on the radiologist/user interaction with the tool; (1) fully automated techniques that require no, or negligible user input, and (2) semi-automated techniques that require some localization or initialization from the user; then the algorithm provides the majority of segmentation optimization. Semi-automated techniques outperform automatic approaches, resulting in sufficiently accurate and robust results (Zhao and Xie, 2013). However, semi-automated techniques do not scale well to large number of labeled datasets, since developing and validating interactive algorithms becomes laborious as the datasets grow. Consequently, there is an unmet need for an approach to simulate user interaction that will allow for efficient and cost-effective evaluation of semi-automated techniques throughout the development and validation stages.

In a recent publication we presented Semi-Automated Map-BAsed Segmentation (SAMBAS), which allows for real-time feedback by an expert radiologist (Gering et al., 2018). In short, the user initializes the segmentation process by drawing a long axis; during the long axis drawing, the 2D segmentation updates in real-time for interactive feedback. In cases of suboptimal 2D segmentation the user can refine the result by drawing a short axis. Further optimization can be performed on the other two planes prior to 3D segmentation initialization. This interactive system outperformed the Deep Learning (DL) approach alone; as demonstrated in our publication, using the Multimodal Brain Segmentation Competition (BraTS) 2018 validation data the interactive system resulted in an improved Dice similarity coefficient over DL alone and the lowest Hausdorff-95% distance on the BraTS leaderboard (Menze et al., 2015; Bakas et al., 2017a, 2018; Gering et al., 2018).

However, it is still unknown how real-time experiential input affects Dice coefficient and Hausdorff-95% distance. Therefore, in this study, we designed an experiment to simulate the level of user interaction. Specifically, we used the 2018 BraTS training data as the ground-truth and a computer simulation mimicked the process that a radiologist would follow to perform segmentation with real-time interaction (Bakas et al., 2018). Clicks and drags were placed only where needed in response to the deviation between real-time segmentation results and assumed radiologist's goal, as provided by the ground-truth. Results of accuracy for various levels of interaction are presented along with estimated elapsed time, in order to measure efficiency.

MATERIALS AND METHODS

Rapid Precise Metrics™ (RPM) implements an interactive algorithm as a probabilistic framework with efficient user interaction and control in the HealthMyne® Platform (HealthMyne, Madison, WI). Additional details on how RPM seamlessly merges DL with user interaction can be found on Gering et al. (2018). For the purposes of this work, we removed the DL component because access was needed to the same ground-truth data on which the DL would have trained. Indications by a skilled radiologist are another aspect of RPM missing from this experiment. Consequently, the absolute values of accuracy reported do not fully reflect clinical performance of RPM, however, relative accuracy and timing measurements should be representative.

The organization of the manuscript is as follows: the system for interactive Multi-Plane Reformat (MPR) segmentation will be described first, followed by the method for simulating a user's interaction with the system. Finally, the method for performing timed tests will be presented.

Interactive Multi-Plane Reformat (MPR) Segmentation

Like a digital simulation of a traditional light box on which radiologists formerly viewed film, the 3D volume is visualized by displaying 2D planes sequentially. A MPR refers to reformatting more than one plane, such that a trio of planes is displayed side-by-side corresponding to axial, coronal, and sagittal orientations (**Figure 1**).

The user initializes the segmentation process by drawing a long axis on one plane of the MPR. As the user draws the long axis, a 2D segmentation updates in real-time for interactive feedback. The feedback has proven to be very helpful for the user to know precisely where to place the endpoint of the axis (Gering et al., 2018). Upon release of the mouse, 2D segmentation occurs immediately on the other MPR planes. **Figure 2** shows the interactive feedback.

When the 2D contour is unsatisfactory, additional drags may be drawn to complement the long axis. Furthermore, single clicks may be used to "drop" points along the structure boundary. Another available editing operation is a "ball tool" for drawing with a digital brush. A correct 2D segmentation is important since probability distributions are learned from the 2D segmentation and employed in segmenting the other MPR planes.

When the contours on other MPR planes are unsatisfactory, then the user can further refine the segmentation by either drawing long axes on these planes, or by editing the segmentation masks with the ball tool. This is especially useful for irregularly shaped lesions or lesions oriented

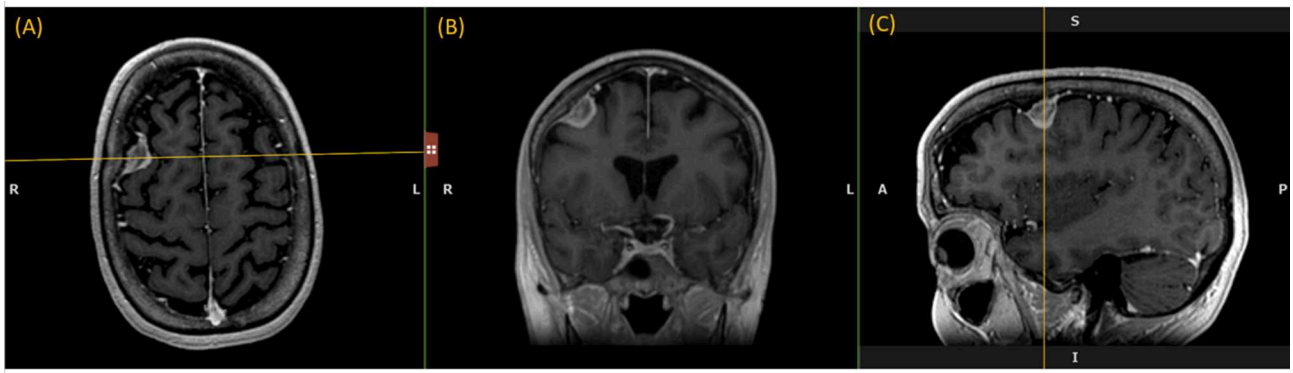


FIGURE 1 | Multi-Plane Reformat: an example of Multi-Plane Reformat (MPR) slices of the 3D volume corresponding to axial, coronal and sagittal planes. The yellow lines denote the position of the coronal plane **(B)**, with respect to the axial plane **(A)** and the sagittal plane **(C)**.

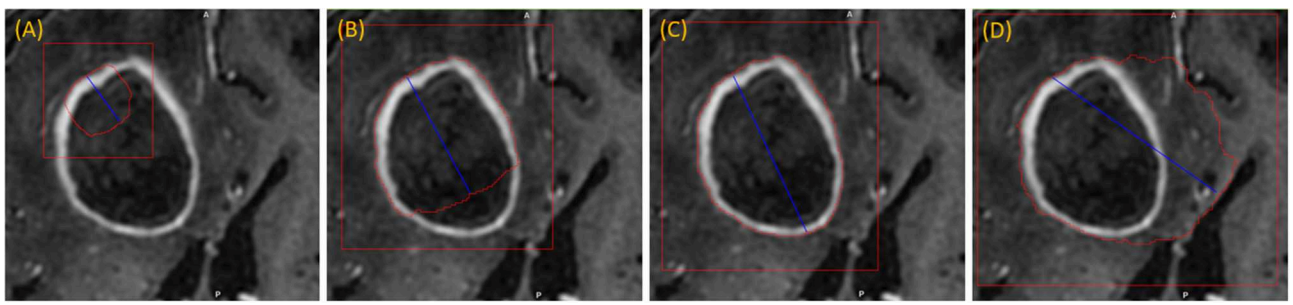


FIGURE 2 | Interactive segmentation: several instances during interactive segmentation **(A–D)** depicting the segmentation contour (red) updating in real-time as the user drags the endpoint of the long axis (blue). In **(C)** a correct segmentation of core tumor is displayed, while **(D)** displays response to overdrawing.

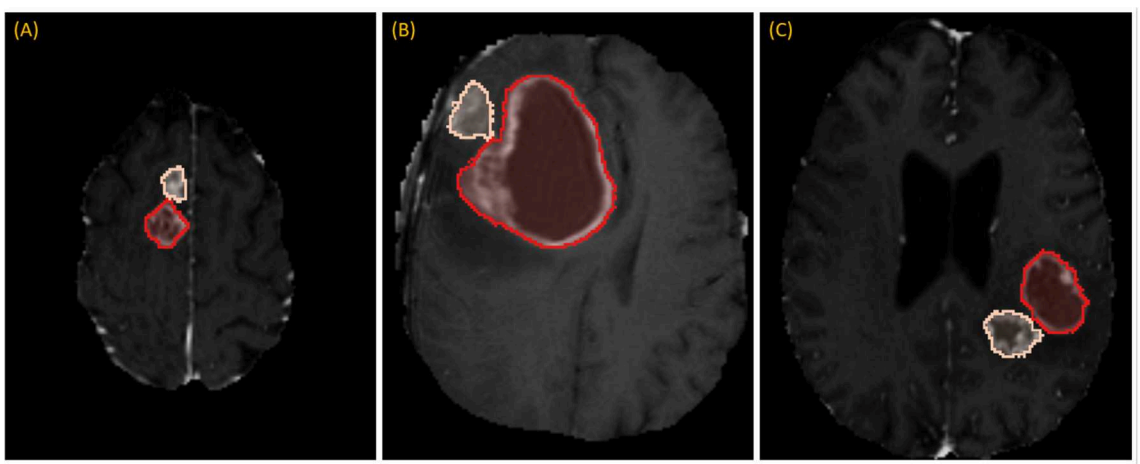


FIGURE 3 | Separating adjoining tumors: three examples of adjoining tumors **(A–C)** which were manually separated to form distinct tumors, shown here in tan and red.

obliquely to the anatomical axes. Once initial segmentation is satisfactory, the user can initiate 3D segmentation by a single click.

3D segmentation occurs quickly (approximate time = 1–2 s), and the user may inspect the resulting contours by scrolling through slices on any MPR plane. If unsatisfied, the user has two options, either delete the lesion segmentation and re-draw a better long axis, or alternatively edit the 3D segmentation using a 3D sphere tool. When satisfied, the user clicks a button to confirm the 3D contours.

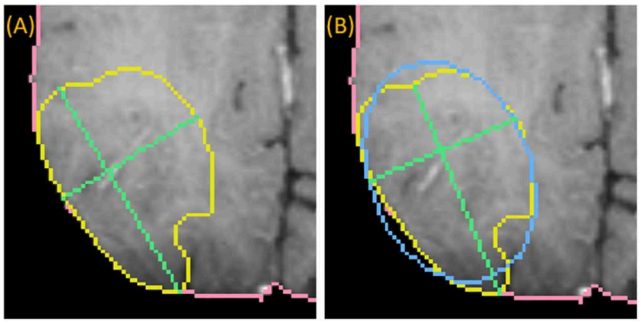


FIGURE 4 | Medially placed longest axis: a demonstration of a more medially placed longest axis obtained by fitting an ellipse (blue) to the ground-truth (yellow) and finding the long and short axes (green) parallel to the major and minor axes of the ellipse. The true longest axis is drawn in **(A)**, while the more medially positioned axis is drawn in **(B)**.

Simulation System

The simulation system automatically draws a long axis on each tumor. Depending on the accuracy of the resulting segmentation, more drags or clicks are added as needed. The process by which these are drawn aims to mimic a human user's actions, as described below.

Data Preparation

Multi-institutional, routine clinically-acquired, pre-operative, multispectral MR scans were provided by the 2018 BraTS challenge (Bakas et al., 2017b,c, 2018). The data have been preprocessed to be co-registered to the same anatomical template, interpolated to the same resolution (1 mm^3), and skull-stripped.

For the purposes of this work, we used only the post-contrast T1-weighted MR scans. While BraTS provided labels for three phenotypes: whole tumor, core tumor, and active tumor, we combined the ground-truth masks for "core" and "active" to form "core tumor," and used this one tumor component exclusively in this experiment, since this is representative of the gross tumor extent assessed by a radiologist in the clinical setting.

Enumerating Tumors

While RPM is designed to segment individual lesions, BraTS ground-truth presents a unified mask without separating individual lesions. Therefore, we manually drew blank (zero-valued) lines to separate adjoining lesions. After this one manual step, lesions could be enumerated automatically by running 3D connected-component analysis (CCA) to identify each distinct "island" of the ground-truth mask, as illustrated in **Figure 3**. Given the 285 patients, 232 had 1 lesion, 33 had 2 lesions, and 20 had 3 or more, with the maximum being 5.

Simulating the Drawing of a Long Axis

The first step toward drawing a long axis is selecting the axial slice on which to draw. Our aim was to replicate the approach of an expert radiologist briefly scrolling through the slices to eyeball the one on which the tumor appears the largest. In the first step, for each enumerated tumor, the range of slices containing ground-truth was found, and the subset of slices in the central third was considered. Given this subset, the slice with the largest area of ground-truth was chosen.

In order to simulate the type of long axis that a user might draw, we employed four different methods: (i) identification of the true longest axis, (ii) selection of an axis that is located more medially than the true longest axis, (iii) search for an axis that includes pixels that statistically typify the tumor, and (iv) sweeping a short distance to search for optimal results.

To draw the "medial" long axis, an ellipse is fit by Principle Component Analysis (PCA) to the 2D segmentation (Duda et al., 2012). The long axis with the same orientation as the major axis of the ellipse is selected, as shown in **Figure 4**. This method is driven by the fact that RPM tends to perform better on centrally located drags where symmetry can be exploited.

Since RPM samples statistics under the long axis, it's important to consider that aspect in addition to length and centrality. Therefore, the third method searches the set of N^2 possible axes drawn between a set of N points spaced near each endpoint of the medial axis. Based on the ellipse, these points are spaced by a few degrees, and lie on the boundary of the ground-truth. A score is computed for each axis, and the axis with the best score is selected. Equation (1) describes the score as a weighted combination of properties of the long axis, namely length, centrality, and relative entropy, also referred to as Kullback-Leibler divergence (D_{KL}) (Cover and Thomas, 2012).

$$Score = \alpha * (1 - D_{KL}) + \beta * Length + \gamma * Centrality \quad (1)$$

where α, β and γ are scalar parameters. Since the Kullback-Leibler divergence is a measure of how one probability distribution is different from another, it is an appropriate metric for evaluating how well the pixels along the long axis relate to the pixels of the entire 2D structure. Equation (2) expresses this relationship where the probability distribution, Q, of pixels sampled under the long axis is estimated by Parzen window density estimation, and the probability distribution, P, is estimated similarly from pixels sampled under the ground-truth mask (Duda et al., 2012).

$$D_{KL}(P||Q) = -\sum_{x \in X} P(x) \log\left(\frac{Q(x)}{P(x)}\right) \quad (2)$$

The *Length* and *Centrality* in Equation (1) are terms with a range [0, 1] and are computed as follows:

$$Length = \frac{length\ of\ axis}{length\ of\ true\ longest\ axis} \quad (3)$$

$$Centrality = 0.1 + 0.9 * C/(2r) \quad (4)$$

Where C encodes the distance from center of ellipse obtained for Equation (5):

$$C = |i - r| + |j - r| \quad (5)$$

Where indices, i and j, index the sets of N points on each side of the axis, and r represents the index, $N/2$, of the middle point in

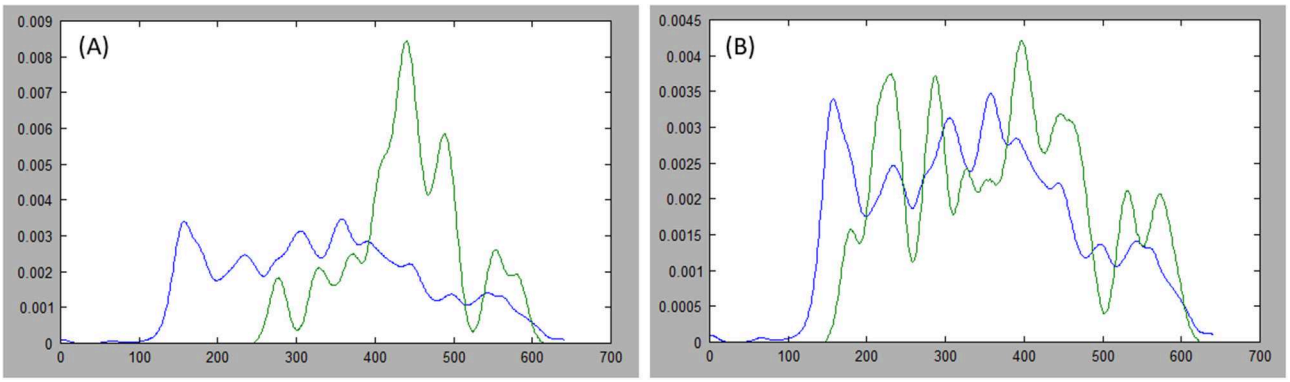


FIGURE 5 | Comparing distributions: estimations of the probability distribution function (PDF) for tumor ground truth and pixels under the long axis are plotted. A high discrepancy is plotted in **(A)** resulting in large Kullback-Leibler (KL) divergence, while the PDFs are more similar in **(B)** resulting in small KL divergence.

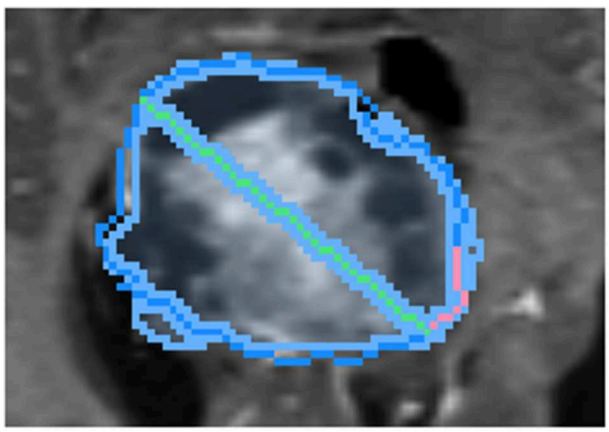


FIGURE 6 | Demonstration of sweeping approach for long axis identification: while the leftmost endpoint of the long axis (green) is held fixed, the rightmost endpoint is swept along a short path (pink) along the boundary of ground-truth (dark blue). At each position, the similarity between the segmentation (light blue) and ground-truth is measured.

a set. Furthermore, we favor axes that cross the center by halving the expression above when i and j both lie on the same side of r. **Figure 5** presents examples of probability distribution functions for a long axis that is representative of the tumor, and another axis that is divergent.

To draw the "swept" long axis, the first endpoint of the long axis is held fixed while the second endpoint is dragged along the boundary of the structure, by a short distance in each direction, as shown in **Figure 6**. At each position during the sweep, the interactive 2D segmentation is performed, and the position with the best comparison with ground-truth is selected (by DSC defined below).

Simulating the Drawing of a Short Axis

While the RPM algorithm allows multiple drags of any orientation, we simplified this experiment by drawing only the "short axis," which is defined as the longest axis that lies perpendicular to the long axis.

Simulating Dropped Points Along Structure Boundary

Given a segmentation based on the long and short axes, the contour point of greatest disagreement with the ground-truth is identified. Subsequently, an editing operation is performed by "dropping" a point on the structure boundary as indicated by the ground-truth. As described earlier, these drops serve as inputs into RPM's algorithm that are quicker to draw than a line with two endpoints.

Following the first dropped point, segmentation is recomputed and the next contour point of greatest disagreement is identified, if any, as there may be no remaining significant discrepancies. New points cannot be placed too closely to earlier points. In this manner, more points can be "dropped" in succession, triggering new segmentations with each dropped point (**Figure 7**).

Simulating Drawing on MPR

The center of the long axis is used to determine the center of the reformatted sagittal and coronal planes that comprise the 3-plane MPR. Long and short axes were drawn in similar manner on all planes. The additional axes precipitate MPR segmentation.

Comparison of Volumes

While each tumor is segmented individually, the "ground-truth" is provided per patient rather than per lesion. Consequently, we took the union of the segmentations of all lesions for each patient, and compared these aggregates with the "ground-truth" for agreement. The Dice similarity coefficient (DSC) was used to measure the similarity between two sets of segmentations and was calculated using the Equation (3):

$$DSC = \frac{2(A \cap B)}{(A + B)} \tag{6}$$

where A represents the semi-automated segmentation and B represents the "ground-truth" (Allozi et al., 2010). Scores were

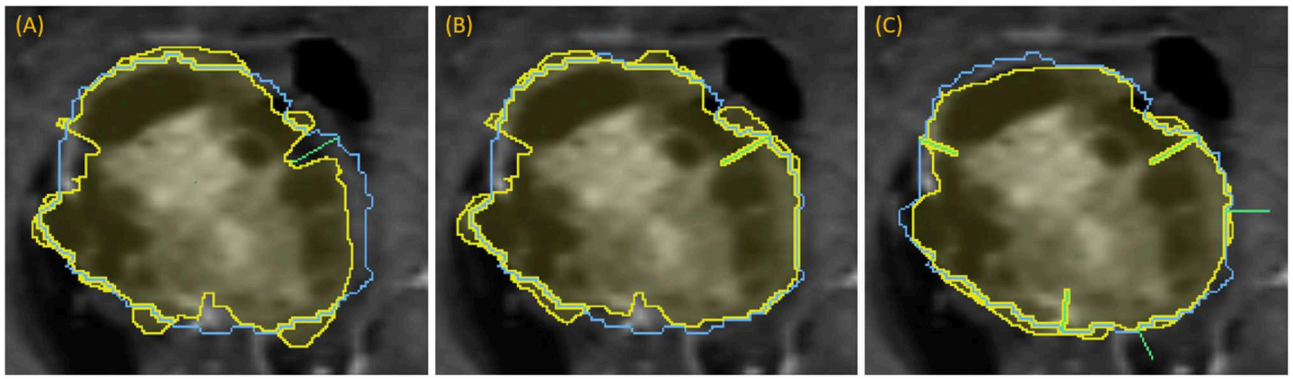


FIGURE 7 | Demonstration of dropping points: segmentation (yellow) relative to the ground-truth outline (blue) for a progression of dropping points to segment a complex lesion: **(A)** green line denotes where a dropped point is most needed, **(B)** updated segmentation after dropping that point, **(C)** updated segmentation after dropping five points, each indicated with a green line.

TABLE 1 | Timed tasks.

Load study
Scroll to lesion
Segment by dragging long axis
Optionally open MPR view for additional drags/clicks
Perform 3D segmentation
Scroll to inspect 3D contours
Confirm 3D contours

computed by uploading segmentations to the CBICA Image Processing Portal.

Timing Tests

To estimate an average elapsed time, the entire population was partitioned into three categories so that a weighted average could be computed where the weights are determined based on the size of each category. From each category, 10 cases were randomly sampled (by Python script) to be segmented by a human user with the interactive system; the user has extensive background in the design and implementation of clinical software solutions, though no specific radiology training. The total elapsed time measured included all of the tasks listed in **Table 1**, which span from beginning to load the study to confirming 3D contours.

The three categories were the following: (i) cases that segmented well with drawing an axial long axis (DSC $>$ 0.883; $n = 109$); (ii) cases that required drawing long axis on all MPR planes ($n = 123$); (iii) cases that required additional edits ($n = 53$). Prior to the random selection, the categories were whittled down in size by more restrictive criteria in order to form subsets of patients for which the interaction was more meaningful. These subsets were cases that segmented extremely well (DSC $>= 0.93$) with a single drag ($n = 39$), cases whose score increased by at least 0.05 to achieve a total score of at least 0.85 given a long axis drag on all MPR planes ($n = 25$), and cases whose score increased further by at least 0.05 to achieve a total score of at least 0.85 given additional edits on MPR ($n = 41$).

TABLE 2 | Comparison of various strategies for simulating the drawing of the long axis.

Long axis style	3D dice
True longest	
Mean	0.798
St. Deviation	0.130
Range	[0.315–0.972]
Favoring medial position	
Mean	0.812
St. Deviation	0.122
Range	[0.232–0.963]
Searching for statistics	
Mean	0.807
St. Deviation	0.130
Range	[0.232–0.963]
Sweeping one endpoint	
Mean	0.821
St. Deviation	0.120
Range	[0.305–0.972]

To evaluate the realistic nature of simulations against real user interaction, DSC score from manual segmentation was compared against DSC obtained from simulations. First, the fitness of DSC scores to normal distribution was determined by Kolmogorov-Smirnov test, and then the scores were compared using t-test or Mann-Whitney test accordingly. A p-value of <0.05 was considered statistically significant. Finally, for completeness we calculated the DSC score between the segmentation obtained from real user interaction and from simulation.

RESULTS

Of the 285 patients, 232 had a solitary lesion, and 53 had more than one lesion resulting in a total of 365 brain lesions available for segmentation.

Long Axis Simulation

We simulated four different strategies for drawing the long axis: (i) obtaining the true longest axis, (ii) assigning the long axis more medially than the true longest axis, (iii) searching for an axis that statistically typifies the tumor, and (iv) sweeping a short distance to search for optimal results. Using the aforementioned strategies as initialization step, the 3D segmentations were compared with the "ground-truth." **Table 2** summarizes the results for DSC between the four strategies, with swept being noticeably superior (DSC = 0.821).

TABLE 3 | Varying levels of interaction.

User input	Axial plane only	All 3 MPR planes
Long axis only		
Mean	0.821	0.851
St. Deviation	0.120	0.079
Range	[0.305–0.972]	[0.517–0.965]
Long and short axes		
Mean	0.823	0.858
St. Deviation	0.110	0.068
Range	[0.385–0.973]	[0.512–0.965]
Long and short axes and few dropped points		
Mean	0.834	0.864
St. Deviation	0.103	0.063
Range	[0.309–0.973]	[0.557–0.965]
Edited to perfection		
Mean	0.839	0.890
St. Deviation	0.105	0.050
Range	[0.431–0.970]	[0.681–0.970]

TABLE 4 | Progression through processing stages.

Stage	True long	Sweep long	Long and short	Long, short and drops	Axial perfect	MPR perfect
2D	0.886	0.916	0.919	0.947	0.972	0.968
3-plane Scout	0.845	0.859	0.866	0.878	0.886	0.960
3D	0.798	0.821	0.823	0.834	0.839	0.890

Simulating User Interaction

We simulated a varied degree of user interaction from drawing only one axis to editing 3D segmentation to perfection. Additionally, simulations allowed for drawing on the axial plane only, or on all 3 MPR planes: axial, coronal, and sagittal. **Table 3** summarizes the results of 3D segmentations with varying degrees of user interaction. Our results indicate that drawing long axes on MPR planes compared to drawing only one long axis on the axial plane resulted in significantly (by at least 0.05) improved DSC scores in 76 patients out of a total of 285. Similarly, drawing short axes and dropping points on MPR planes resulted in significantly higher DSC scores in 88 patients, while in 14 patients the DSC score worsened significantly. Finally, editing segmentation outcome to perfection on MPR planes significantly improved the DSC score in 123 patients while significantly worsening only 2.

Table 4 presents results from a few intermediate stages of the algorithm. For responsive interaction, RPM segments first in 2D, corresponding to the first row of **Table 4**, and then it initializes the 3D segmentation by segmenting on a 3-plane "scout" reformat, corresponding to the second row of the **Table 4**, and then it finally segments in 3D. There is a column for each level of user interaction.

Timing Tests

The average elapsed time for each patient in the entire population was estimated to be 46.2 s; this was a weighted average computed over three categories whose boundaries were described in section Timing Tests. The number of cases in the first category (drawing an axial long axis) is the total number of cases whose scores were above 0.883, which was 109. The number of cases in the second category (drawing long axes on all MPR planes) was the number

TABLE 5 | Timing measurements.

User input	Mean elapsed time (seconds)	DSC (user vs. ground-truth)	DSC (simulation vs. ground-truth)	*p*-value	DSC (user vs. simulation)
Long axis only					
Mean	30.38	0.934	0.943	0.1041	0.951
St. Deviation	6.41	0.013	0.008		0.021
Range	[23.65–44.92]	[0.911–0.952]	[0.933–0.954]		[0.918–0.985]
Long axis on 3 MPR planes					
Mean	52.0	0.882	0.876	0.3075	0.877
St. Deviation	17.70	0.063	0.035		0.059
Range	[31.75–86.61]	[0.719–0.935]	[0.8162–0.928]		[0.792–0.962]
Long and short and few dropped points, on MPR					
Mean	65.31	0.844	0.857	0.5205	0.825
St. Deviation	18.01	0.054	0.029		0.063
Range	[46.73–107.75]	[0.729–0.923]	[0.817–0.907]		[0.718–0.921]

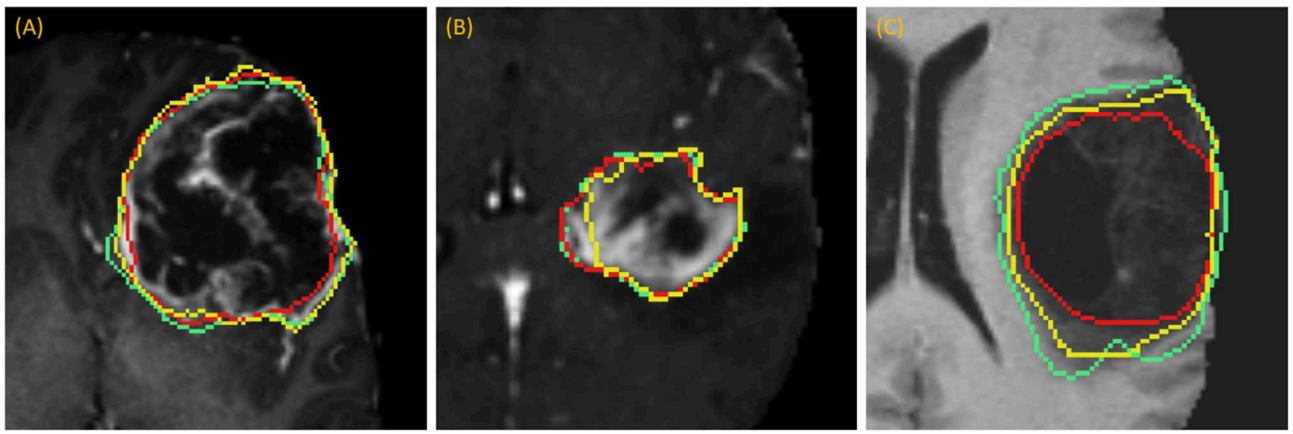


FIGURE 8 | Accuracy Comparison: segmentation outlines by simulation (green), user (red) relative to the ground-truth (yellow) for one case in each of the three categories **(A)** drawing an axial long axis, **(B)** drawing long axes on all MPR planes, and **(C)** performing additional edits required by dropping points.

of the remaining patients whose scores were at least 0.796 given MPR drags, which is 123. Finally, 53 cases comprised the third category (additional edits required).

Table 5 summarizes the results of measuring elapsed time for a user to segment a batch of 10 cases from each of the three categories of interaction. Regarding the first category where the long axis is drawn only on the axial plane, the average time the segmenter reported was 30.38 s, ranging from 23.65 to 44.92 s. One of the 10 cases was large and highly heterogeneous, and therefore the user had more slices to sort through to determine a good place to drag; it should be noted that the initial drag for this case was deleted and redrawn from a better angle. Similarly, drawing the long axis on all three MPR planes resulted in an average time of 52.0 s (range: 31.75–86.61 s). In the final category, dragging lines and dropping points, as well as drawing using the ball tool, as needed on MPR, resulted in an average time of 65.31 s (range: 46.73–107.75 s).

DSC scores obtained from manual segmentation and simulations are presented in **Table 5**. Simulated interaction performed marginally better in the first and third categories, while real interaction scored moderately better in the second category, though none of those differences were statistically significant. Further, DSC scores obtained by comparing user segmentation and simulation are in the same range as DSC scores with ground-truth. Given the three levels of user interaction, segmentation results are depicted in **Figure 8**. In most cases, the "between" DSC score was higher than both individual scores computed relative to ground-truth (**Figure 8B**). In **Figure 8A**, the "between" score is between the individual scores, which in this case, occurs because the user under-segmented the tumor area. Lastly, **Figure 8C** illustrates the third pattern of the "between" score lying below both individual scores. In this case the user under-segmented while the simulation over-segmented, resulting in great disparity between the two. Note, in the case depicted in **Figure 8C** there is no enhancing component surrounding the necrotic core, a typical presentation of a GBM, which would aid identification of tumor border and increase agreement between user and simulation.

DISCUSSION

As large-scale labeled image datasets are being curated for academic challenges and training DL models, the most common application tends to be the development of fully automatic methods for tumor segmentation that do not involve user interaction. One driver of this trend might be that developing and validating interactive algorithms becomes laborious as the datasets grow, owing to the time required to interact with every case in the validation dataset. Further, validation ideally occurs very frequently, interspersed between algorithm updates, and throughout the process of algorithm development. However, for a segmentation algorithm to be clinically applicable its outcome should be optimal, i.e., similar to the ground-truth, and therefore it's expected to require "some" user input (Langlotz et al., 2019). Our research aims to provide an approach to automate many aspects of user interaction and thus expedite large-scale validation.

Given a labeled dataset, it seems natural to employ it for validation by measuring the true longest axis, and then using that as an input to an interactive algorithm such as RPM. However, our results showed that the true longest axis would be a poor choice, as it scored the lowest of the four strategies listed in **Table 2**, where "sweeping" the long axis proved the optimal approach.

Further, our results showed a steady improvement in 3D accuracy as interaction increases on axial plane; DSC scores increased from 0.82 to 0.89. We demonstrate that drawing on MPR markedly improves accuracy vs. drawing on axial planes alone. Somewhat surprisingly, drawing a short axis in addition to the long axis made a rather insignificant improvement in DSC. Motivated by this finding, we changed RPM's design to accept multiple arbitrary lines rather than a single line constrained to be perpendicular to the long axis. Therefore, the user may draw the line through image content whose brightness needs to be sampled in order to complement the sampling already performed by the long axis.

Future work will be inspired by the results of **Table 4**, which suggest that accuracy in the 2D segmentations falls off moderately as the algorithm advances to the scout segmentations, and does so again during advancement to the 3D segmentation. Although each stage of the algorithm performs some machine learning to glean information from the results of the prior stage, perhaps more can be done in this regard. Future work will also investigate why additional editing operations worsened scores for certain patients as the statistical sampling from the initial long axis appears to have been better suited for application to 3D segmentation.

Table 5 reveals that accuracy was comparable between the simulated interaction and the real human interaction. Simulated interaction performed marginally better in the first and third categories, while real interaction scored moderately better in the second category. Human interaction scores lower when the segmenter and creators of ground-truth have a difference of opinion. This effect is somewhat canceled out by the fact that human interaction scores higher when the segmenter can apply more intelligence than that embodied by the simulation algorithm.

Timing results were extremely fast when comparing with the limited number of reports currently published; one study measured lung lesion contouring to require an average of 10.31 min (Velazquez et al., 2013). Our sub-minute timing confirms that RPM enables segmentation in routine clinical use.

For a technical description of how the algorithm compares with popular interactive methods, the reader is referred to Gering et al. (2018). Recent works have also introduced simulated user interaction to either identify the object and initiate segmentation (Xu et al., 2016) or refine the segmentation output obtained by DL (Wang et al., 2018; Zhou et al., 2019). In contrast, our goal in this manuscript was to simulate the approach a radiologist would follow for initiating a segmentation and providing input real time until the optimal result is achieved. When comparing the approach Xu et al. followed for object identification, their mode of interaction was to drop points, in contrast to our mixture of lines and points (Xu et al., 2016). Further, their objective was to generate tens of thousands of training samples and points are sampled randomly from the foreground and background object interiors with spacing constraints (Xu et al., 2016). For comparison with other methods which accept user strokes instead of just points, these randomly sampled points are expanded to circles of radius 5 pixels. Observe that human users would draw free-form strokes rather than perfect circles, so our method differs by its intention to more realistically mimic the actions of a human user, and by its purpose of facilitating continuous algorithm development. Last but not least, in our work we performed direct comparison of the outcome obtained from simulations with the outcome obtained by a human user to demonstrate that our simulations realistically capture real user interaction (**Table 5** and **Figure 8**).

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

DG conceived the idea, developed the theory, and performed the computations. BY-M performed the user segmentations for the Timing Tests. DG contributed to the interpretation of the results. DG and AK wrote the manuscript. NM, AA, and TM provided critical feedback during manuscript preparation. LK, HK, JH, RC, and LP provided critical feedback in RPM discussions and design sessions. All authors approved the final manuscript.

FUNDING

This study was partially supported by NSF #1345927 (RC).

REFERENCES

Allozi, R., Li, X. A., White, J., Apte, A., Tai, A., Michalski, J. M., et al. (2010). Tools for consensus analysis of experts' contours for radiotherapy structure definitions. *Radiother. Oncol.* 97, 572–578. doi: 10.1016/j.radonc.2010.06.009

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. (2017c). *Segmentation Labeld and Radiomic Features for the Pre-Operative Scans of the TCGA-LGG Collection*. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). *Segmentation Labels and Radiomic Features for the Pre-Operative Scans of the TCGA-GBM Collection*. The Cancer Imaging Archive.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:* 181102629.

Cover, T. M., and Thomas, J. A. (2012). *Elements of Information Theory*. New York, NY: John Wiley & Sons.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern Classification*. New York, NY: John Wiley & Sons.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* 45, 228–247.

Gering, D., Sun, K., Avery, A., Chylla, R., Vivekanandan, A., Kohli, L., et al. (2018). "Semi-automatic brain tumor segmentation by drawing long axes on multi-plane reformat," in *International MICCAI Brainlesion Workshop* (Berlin: Springer), 441–455.

Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., et al. (2019). A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/the academy workshop. *Radiology* 291, 781–791. doi: 10.1148/radiol.2019190613

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Omuro, A., and DeAngelis, L. M. (2013). Glioblastoma and other malignant gliomas: a clinical review. *JAMA*. 310, 1842–1850. doi: 10.1001/jama.2013.280319

Sorensen, A. G., Batchelor, T. T., Wen, P. Y., Zhang, W. T., and Jain, R. K. (2008). Response criteria for glioma. *Nat. Clin. Pract. Oncol.* 5, 634–644. doi: 10.1038/ncponc1204

Velazquez, E. R., Parmar, C., Jermoumi, M., Mak, R. H., van Baardwijk, A., Fennessy, F. M., et al. (2013). Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci. Rep.* 3:3529. doi: 10.1038/srep03529

Wang, G., Zuluaga, M. A., Li, W., Pratt, R., Patel, P. A., Aertsen, M., et al. (2018). DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1559–1572. doi: 10.1109/TPAMI.2018.2840695

Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., et al. (2010). Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J. Clin. Oncol.* 28, 1963–1972. doi: 10.1200/JCO.2009.26.3541

Wolchok, J. D., Hoos, A., O'Day, S., Weber, J. S., Hamid, O., Lebb,é, C., et al. (2009). Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. *Clin. Cancer Res.* 15, 7412–7420. doi: 10.1158/1078-0432.CCR-09-1624

Xu, N., Price, B., Cohen, S., Yang, J., and Huang, T. S. (2016). Deep interactive object selection. *Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV).

Zhao, F., and Xie, X. (2013). An overview of interactive medical image segmentation. *Ann. BMVA* 2013, 1–22.

Zhou, B., Chen, L., and Wang, Z. (2019). Interactive deep editing framework for medical image segmentation. *Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen).

Conflict of Interest: DG, AA, and RC are named co-inventors on a patent application (Patent Application No. PCT/US2018/040473. DG, AA, JH, BY-M, LK, HK, RC, and LP are named co-inventors on a patent application (Patent Application No. PCT/US2019/059897). TM declares an equity interest and advisory role to HealthMyne, Inc. DG, BY-M, NM, AA, LK, HK, JH, RC, and LP declare equity interest to HealthMyne, Inc. AK declares salary support from The University of Texas MD Anderson Cancer Center, and HealthMyne, Inc.

frontiers in Neuroscience

METHODS
published: 28 April 2020
doi: 10.3389/fnins.2020.00350

3D-BoxSup: Positive-Unlabeled Learning of Brain Tumor Segmentation Networks From 3D Bounding Boxes

Yanwu Xu[1], *Mingming Gong*[2], *Junxiang Chen*[1], *Ziye Chen*[3] *and Kayhan Batmanghelich*[1]*

[1] *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States,* [2] *School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC, Australia,* [3] *School of Compuer Science, Wuhan University, Wuhan, China*

Accurate segmentation is an essential task when working with medical images. Recently, deep convolutional neural networks achieved a state-of-the-art performance for many segmentation benchmarks. Regardless of the network architecture, the deep learning-based segmentation methods view the segmentation problem as a supervised task that requires a relatively large number of annotated images. Acquiring a large number of annotated medical images is time consuming, and high-quality segmented images (i.e., strong labels) crafted by human experts are expensive. In this paper, we have proposed a method that achieves competitive accuracy from a "weakly annotated" image where the weak annotation is obtained via a 3D bounding box denoting an object of interest. Our method, called "3D-BoxSup," employs a positive-unlabeled learning framework to learn segmentation masks from 3D bounding boxes. Specially, we consider the pixels outside of the bounding box as positively labeled data and the pixels inside the bounding box as unlabeled data. Our method can suppress the negative effects of pixels residing between the true segmentation mask and the 3D bounding box and produce accurate segmentation masks. We applied our method to segment a brain tumor. The experimental results on the BraTS 2017 dataset (Menze et al., 2015; Bakas et al., 2017a,b,c) have demonstrated the effectiveness of our method.

Keywords: brain tumor segmentation, deep learning, weakly-supervised, 3D bounding box, positive-unlabeled learning

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Shijie Zhao,
Northwestern Polytechnical University,
China
Madhura Ingalhalikar,
Symbiosis International University,
India

****Correspondence:***
Kayhan Batmanghelich
kayhan@pitt.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *31 August 2019*
Accepted: *23 March 2020*
Published: *28 April 2020*

Citation:
Xu Y, Gong M, Chen J, Chen Z and Batmanghelich K (2020) 3D-BoxSup: Positive-Unlabeled Learning of Brain Tumor Segmentation Networks From 3D Bounding Boxes. Front. Neurosci. 14:350. doi: 10.3389/fnins.2020.00350

1. INTRODUCTION

Gliomas are one of the most common brain tumors in adults. They can be categorized into different levels of aggressiveness, including High-Grade Gliomas (HGG) and Lower Grade Gliomas (LGG) (Louis et al., 2016). Gliomas consist of heterogeneous histological sub-regions, including peritumoral edema, the necrotic core, as well as the enhancing and non-enhancing tumor core (Menze et al., 2015). Magnetic Resonance Imaging (MRI) of brain tumors is commonly used to evaluate tumor progression and plan treatments. An MRI usually contains multi-modal data, such as T1-weighted, T2-weighted, contrast enhanced T1-weighted (T1ce), and Fluid Attenuation Inversion Recovery (FLAIR) images, which provide complementary information for analysis of brain tumors.

The automatic segmentation of brain tumors and subregions is a crucial pre-treatment step for the characterization and subtyping ofgliomas. This is a challenging problem because tumors vary in shape and size across patients and may have low contrast in some modalities. Recently, deep convolutional neural network (CNN)-based methods have achieved new records in brain tumor segmentation. Most of these methods are extensions of the U-Net structure (Ronneberger et al., 2015; Çiçek et al., 2016) in various ways (Isensee et al., 2017, 2018; Kamnitsas et al., 2017a,b; Wang et al., 2017; Li et al., 2018). For example, some works focus on the design of new convolutional network structures, such as using a mix between convolutional kernels and modifying the down-sampling strategy (Havaei et al., 2015; Kamnitsas et al., 2017b). Other works have aimed to improve the method of fusing multi-modal information. For example, Wang et al. (2017) suggested a patch-based framework combined with multi-view fusion techniques to reduce false positive segmentation. Kamnitsas et al. (2017a) proposed another fusion method through aggregation of predictions from a wide range of methods. The overall approach is more robust and reduces the risk of over-fitting to a particular dataset.

A key problem of CNN-based segmentation methods is the requirement of accurate pixel/voxel-level annotations. However, annotating a 3D image at the voxel level requires human expertise and is expensive and time consuming. Motivated by a recent work in weakly supervised segmentation in natural 2D images (Dai et al., 2015), we proposed to learn the segmentation network from 3D bounding box annotations. As pointed out in Dai et al. (2015), boxing out the object location is about 15 times faster than drawing the segmented mask (Dai et al., 2015). In 3D MRI images, the burden of annotating voxels is much higher than that of annotating 2D images because the number of voxels increases exponentially with image dimension. However, the cost of 3D bounding box annotations is comparable to that of 2D bounding boxes. Therefore, learning from 3D bounding boxes is valuable for brain tumor segmentation.

In this paper, we have investigated how to train a segmentation network from coarse but easily accessible 3D bounding box annotation. The main difficulty comes from the inaccurate annotations inside the bounding box. More specifically, the region bounded by a 3D bounding box contains tumor voxels as well background voxels. If one simply considers the voxels inside and outside of the bounding box as two classes, i.e., tumor and non-tumor; the non-tumor voxels inside the bounding box will have the wrong labels, and the learned network tends to classify the voxels outside but close to the tumor boundaries as tumor voxels. To solve this problem, we considered segmentation from 3D bounding boxes as a positive-unlabeled (PU) learning problem (Denis, 1998; Elkan and Noto, 2008) in which we consider the voxels outside of the bounding boxes as a positive class and the voxels inside the bounding box as unlabeled data. We have proposed the "3D-BoxSup" method to train a deep convolutional neural network reliably from 3D bounding box annotations with a non-negative risk estimator that is robust against overfitting (Kiryo et al., 2017). We conducted experiments on the BraTS 2017 dataset, and the results show that our method can obtain competitive accuracy by just learning from coarse bounding box annotations.

2. METHODS

Our 3D-BoxSup method is inspired by the BoxSup method (Dai et al., 2015), which aims to segment objects from 2D bounding box annotations. BoxSup is a straightforward method to train deep CNNs from coarse box annotations. It provides a biased objective function and utilizes the updated network in turn to improve the estimated segmentation masks used for training, which means the estimated segmentation masks in the previous training epoch are used as the ground truth mask for the next epoch. However, this iterative method is not practical for the 3D patch-based method because re-calculating the segmentation mask for each volume for each epoch has a high time-cost. In addition, it is unclear whether the iterative method will finally converge to the optimal solution. To achieve a considerable performance without iteratively updating the segmentation masks, we have cast the segmentation problem as a PU learning problem and applied a non-negative PU risk estimator (Kiryo et al., 2017) as the train objective to learn the segmentation network, where we viewed the 3D box annotated region as the unlabeled data and the area outside the box as positive data. In the following section, we have outlined the base model with a biased box-learning estimator as our baseline and the unbiased-box learning method as our proposed method.

In this section, we have first introduced the basic problem setup with precise mathematical definitions. Second, we have introduced our baseline convolutional neural network architecture, used for predicting the segmentation masks. Third, we have presented how the network is usually trained if the ground truth segmentation mask is available. Last, we have described our PU learning-based 3D-BoxSup method and the corresponding algorithm.

2.1. Problem Setup

In the real application, the accurate segmentation mask is difficult to acquire. Thus, in this paper, we only considered the cases where we only had access to box-labeled segmentation data.

Let S_1 denote the annotated 3D box region where gliomas reside and S_0 denote the background area outside of the bounding box. Assuming we extract 3D patches from S_1 and S_0 for training, which is shown in **Figure 1**, the label of each voxel in S_1 and S_0 is assigned to 1 and 0, respectively. The proportion of non-tumor voxels is denoted as $\pi_P = \frac{n_0}{n_0+n_1}$, where n_1 is the number of voxels inside the box and n_0 is the number of voxels outside of the box.

2.2. Training Data Generation

First, for each volume of a patient, we generated a 3D bounding box to roughly cover the whole tumor (WT) area (S_1), and the uncovered region was considered to be the background area S_0. We show an example of the box-labeled data in **Figure 1B**. For convenience, we only showed the box label with a 2D yellow rectangle. It should be noted that the training label in our case is actually a 3D bounding box for each volume. In our experiment,

we generated this 3D bounding box from the accurate ground truth segmentation mask, and we assumed that we did not have access to this accurate segmentation data during the training time, which was available in our testing.

We then followed the standard preprocessing step to process the original 3D input images (Bakas et al., 2018). To reduce the sensitivity to absolute pixel intensities variations, an intensity normalization step is applied to each volume of all subjects by subtracting the mean and dividing by the standard deviation so that each MR volume will have a zero mean and unit variance, which is operated to each volume dependently. In practice, as only the central region that contains the brain is used, the mean and standard deviation are estimated using this brain area; where we exclude the black area outside the brain with voxel value 0. Finally, we extract 200 patches per patient with patch size $48 \times 48 \times 48$ in S_1 and S_0. If the extracted patch of S_1 is partial beyond the boxed area, we pad 0 value to the exceeded part for the segmentation mask. In our experiment, we randomly selected 3D patches from area S_1 and S_0 with a proportion of 0.8 and 0.2, respectively. In all of the following settings, we allocated patches from S_1 with the label 1 and patches from S_0 with the label 0.

2.3. Network Structure

To build a deep network for 3D patch segmentation, we applied the 3D U-Net (Çiçek et al., 2016), consisting of an encoder and a decoder network with skip connections similar to our base model. In contrast to (Çiçek et al., 2016), we removed the last down-sampling layer and the first up-sampling layer for the LHS and RHS of the 3D U-Net, respectively. This is because the down-sampling structure would eliminate edge features of brain tumor. Our modified 3D U-Net is shown in **Figure 1A**.

2.4. Learning With Ground Truth Mask

In the fully supervised brain tumor segmentation task, accurately annotated masks were provided for training. Assuming the mask prediction function modeled by a CNN is $\hat{\mathbf{y}} = f(\mathbf{x}; \theta) \in \mathbf{x} \in \mathbb{R}^{d \times d \times d}$, where $\mathbf{x} \in \mathbb{R}^{d \times d \times d}$ is a randomly chosen 3D patch U from a patient V, and θ is the global trainable parameter. The ground truth patch tumor mask is $\mathbf{y} \in \mathbb{R}^{d \times d \times d}$. To learn the network parameters, we can apply the sigmoid function to generate probability values and cross-entropy loss function to evaluate voxel-wise prediction error. The objective function for a single value in the predicted mask can be written:

$$L_{mask}(y, \hat{y}) = (1 - y) \cdot \log(\frac{1}{1 + e^{-\hat{y}}}) + y \cdot \log(\frac{1}{1 + e^{\hat{y}}}), \quad (1)$$

where y is a single value in the ground truth mask, and $\hat{y}$ is the corresponding value in the predicted mask. The overall empirical risk $\hat{R}_{mask}$ is a summation of $L_{mask}(y, \hat{y})$ on all voxels in all the 3D patches and can be efficiently minimized by using stochastic gradient descent (SGD) methods.

2.5. Positive-Unlabeled Learning With Box Labeled Data

When the images are only provided with bounding box annotations, it is much more difficult to learn the segmentation network $f(\mathbf{x}; \theta)$ because the voxels inside the box can be classified

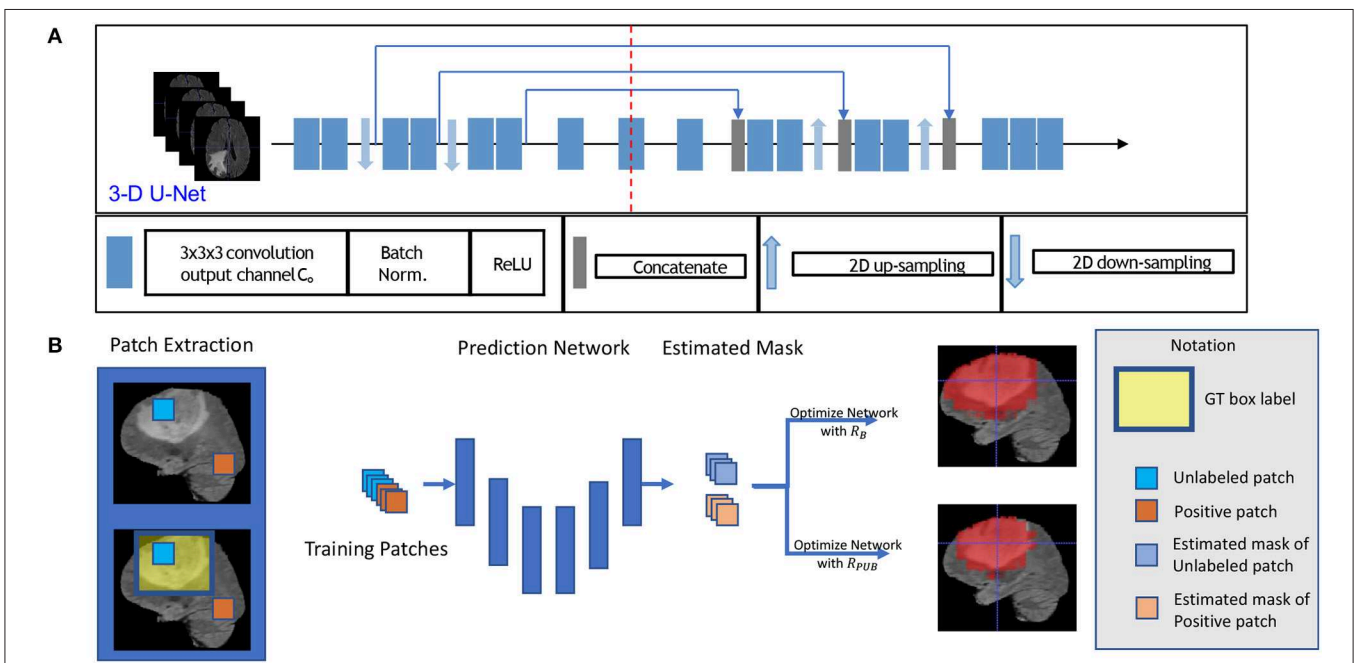


FIGURE 1 | This figure shows the training model with box labeled data. The network structure is shown in sub-figure **(A)**, which is a typical 3D U-Net. The general training process is shown in sub-figure **(B)**. To be intuitive, we applied a 2D slice as our example in **(B)** in the training process, and we fed 3D patches with a 3D bounding box label to optimize our model.

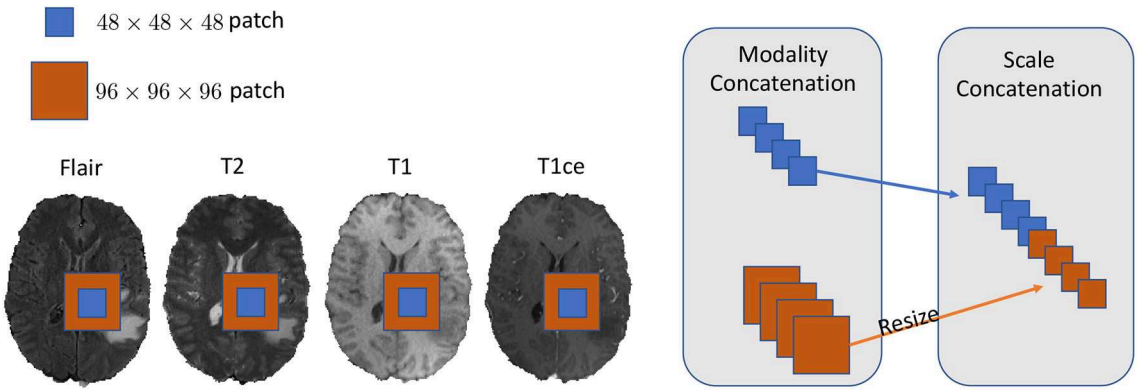


FIGURE 2 | In this figure, we show the example of how to reconstruct our training data in the 2D aspect for better comprehensive as it is easy to implement in 3D level as well. We concatenated the Flair, T2, T1, and T1ce models in the channel dimension and combined extra scale information to enhance the model training.

as either tumor or non-tumor. A straightforward solution would be assigning all the voxels in patches coming from S_1 with label 1 and labeling all the voxels in patches coming from S_0 with label 0. We could then train the segmentation network using the cross-entropy loss (1), which we call the "Naive-BoxSup" method. The problem with the naive method is that some non-tumor voxels inside the box are wrongly assigned with tumor class label 1. As a result, the learned network tends to classify the voxels outside but close to the tumor boundaries as tumor voxels.

To alleviate this problem of the Naive-BoxSup, we proposed to consider segmentation from boxes as a positive-unlabeled learning problem. We can ensure that patches extracted from S_0 only contains positively-labeled voxels (0 is considered the positive label), which are far away from tumor area. In the bounding box area S_1, voxels in S_1 can be considered as an unlabeled object. Thus, segmentation network learning from bounding box annotations is a typical positive-unlabeled learning problem, which tries to learn a classifier to model the distribution of positive data p_p and negative data p_n by using only positive labeled data and unlabeled data. In the following, we have described how we applied a recently proposed non-negative PU-Learning loss (Kiryo et al., 2017) to train our segmentation network. We chose to use this loss because the non-negative constraint on the loss makes it less prone to overfitting when a deep network is being learned.

Let $p(x)$ denote the marginal distribution of input features corresponding to a single output y in the predicted segmentation mask. By stacking all the 3D patches together, we can get a sample $\{(x_i, y_i)\}_{i=1}^{n}$. Let $p_p(x) = p(x|y = 0)$ and $p_n(x) = p(x|y = 1)$ denote the positive and negative class conditional distributions, respectively. We have

$$p(x) = \pi_p p_p(x) + (1 - \pi_p) p_n(x). \tag{2}$$

Equivalently, $(1 - \pi_p)p_n(x) = p(x) - \pi_p p_p(x)$. Let $L(y, \hat{y})$ be a general loss function evaluating the distance between output and ground truth labels, which is cross-entropy loss in our case. This is denoted by

$$R_p^+(\theta) = E_{x\sim p_p(x)} L(f(x,\theta), y = 0), \tag{3}$$

$$R_n^-(\theta) = E_{x\sim p_n(x)} L(f(x,\theta), y = 1), \tag{4}$$

$$R_p^-(\theta) = E_{x\sim p_p(x)} L(f(x,\theta), y = 1), \text{ and} \tag{5}$$

$$R_u^-(\theta) = E_{x\sim p(x)} L(f(x,\theta), y = 1). \tag{6}$$

By using (), we can have an approximation of the risk on the true distribution $R(f) = E_{(x,y)\sim p(x,y)} L(f(x,\theta), y) = \pi_p R_p^+(f) + \pi_n R_n^-(f)$ by

$$R_{PU} = \pi_{\mathrm{p}} R_p^+(\theta) + R_u^-(\theta) - \pi_{\mathrm{p}} R_p^-(\theta). \tag{7}$$

Theoretically, we can minimize R_{PU} to learn the optimal *theta* for our segmentation network. However, as pointed out in Kiryo et al. (2017), if the model is very flexible, empirical risks on training data will go negative, and we will suffer from serious over-fitting. Since our model is a very complicated convolutional neural network, we applied a non-negative risk estimator (Kiryo et al., 2017), as with the objective function:

$$R_{PUB} = \pi_{\mathrm{p}} R_p^+(\theta) + \max\{0, R_u^-(\theta) - \pi_{\mathrm{p}} R_p^-(\theta)\}. \tag{8}$$

In practice, we need to replace the risk terms by their empirical estimates from data:

$$\pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^+(\theta) = -\pi_{\mathrm{p}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i) \cdot \log\left(1 - \frac{1}{1 + e^{-f(x_i;\theta)}}\right)$$

$$\widehat{R}_{\mathrm{u}}^-(\theta) = \frac{1}{n} \sum_{i=1}^{n} y_i \cdot \log\left(\frac{1}{1 + e^{-f(x_i;\theta)}}\right)$$

$$\pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^-(\theta) = -\pi_{\mathrm{p}} \frac{1}{n} \sum_{i=1}^{n} y_i \cdot \log\left(1 - \frac{1}{1 + e^{-f(x_i;\theta)}}\right).$$

The overall algorithm is shown in Algorithm 1. We used the ADAM optimizer to optimize the empirical risk. In our algorithm, we set $\pi_{\mathrm{p}} = 0.75$ and we set $\gamma = 1, \eta = 0.5$, which is a very common choice for PU learning.

3. EXPERIMENT

To demonstrate the effectiveness of our method, we presented a number of experiments examining different aspects of our method. After introducing the implementation details, we evaluated our methods on BraTS (Wang et al., 2017) brain tumor training dataset. We compared the segmentation performance of our 3D-BoxSup method with the Naive-BoxSup segmentation method and show advantages of our proposed method over the baseline approach.

Algorithm 1: Optimization of Our 3D-BoxSup segmentation algorithm

Input: training data (x_i, y_i);
hyperparameters $0 \leq \beta \leq \pi_p$ and $0 \leq \gamma \leq 1$
Output: model parameter θ for $f(x; \theta)$

Let $\mathcal{A}$ be an external ADAM optimizer (Kingma and Ba, 2014)
while no stopping criterion has been met:
  Shuffle (x_i, y_i) into N mini-batches
  for $i = 1$ **to** N:
    if $\widehat{R}_u^-(\theta) - \pi_p \widehat{R}_p^-(\theta) \geq -\beta$:
      Set gradient $\nabla_\theta \widehat{R}_{pu}(\theta)$
      Update θ by $\mathcal{A}$ with its current step size η
    else:
      Set gradient $\nabla_\theta(\pi_p \widehat{R}_p^-(\theta) - \widehat{R}_u^-(\theta))$
      Update θ by $\mathcal{A}$ with a discounted step size $\gamma\eta$

3.1. Training Setting

We got all our training data from BraTS web[1] to evaluate our method. The training data consisted of 285 patients, including segmented masks annotated by human experts. These training data were separated into two categories, including HGG and LGG, each containing 210 HGG and 75 LGG images. There is an imbalance between HGG and LGG, and the data distributions of HGG and LGG were also different, especially for TC and ET. Each patient had four sequences, which are FLAIR, T2, T1, and T1ce. In training time, we randomly split the whole training set to 80% training set and 20% as our evaluation set, and we carried out five folds testing in this manner. We only use the ground truth segmented label during evaluation. We fed all of the sequences into our network by combining them in channel dimension. Thus, our input data are in 5D, the dimensions of which are batch, sequences, width, length, and depth. The training model structure is shown in **Figure 1A**. To generate the training data, we followed the abovementioned section 2.2 method, and the proportion of the non-tumor voxels was $\pi_{S_2} = 0.75$.

We set our training batch size to 64, and each training patch voxel size is $48 \times 48 \times 48$ for saving memory, which is sufficient to train the model; another aspect to the design of such a parch size is that a larger patch would contain more background voxel, which means the model would over-fit the background and would not be able learn a pattern of Whole Tumor segmentation. For the data training strategy, we randomly generated 40,000 locations as the center point of patches for each patient volume; finally, only 200 locations were selected as our training patches. To fully utilize the information from each model provided with FLAIR, T2, T1, and T1ce, we reconstructed our multi-modal data by stack theses modals in the channel dimension, which can be directly fed to convolutional neural networks (CNNs). Also, imitating the technique from (Bakas et al., 2018), we enabled the network to capture the multiscale information from data. To do so, we got the $96 \times 96 \times 96$ patches for each modal, extracted in the same way as the $48 \times 48 \times 48$ patches, which were two times bigger than the basic training patch; this bigger patch also belongs to the same center location as the basic training patch. Then, we resized the $96 \times 96 \times 96$ patches to $48 \times 48 \times 48$. Finally, we concatenated all the different models and scaled patches, which is shown in **Figure 2**. Thus, the input patch size of our model was $batchsize \times 8 \times 48 \times 48 \times 48$. We trained our whole network using Pytorch (Paszke et al., 2017), which is a new hybrid front-end seamlessly transitions between eager mode and graph mode to provide both flexibility and speed. NVIDIA TITAN XP GPU was applied to train our network, and the cost was about 11 gigabytes GPU RAM. The whole training process was finished in 4 h with 5 epochs, and each epoch traversed the whole training dataset. To optimize our model, we chose the common gradient descent algorithm Adam.

3.2. Evaluation Metrics

3.2.1. Dice Coefficient

The Dice-Coefficient Score was calculated as the performance metric. This measure states the similarity between clinical Ground Truth annotations and the output segmentation of the model which are A and B respectively. Afterwards, we calculated the average of those results to obtain the overall dice coefficient of the models.

$$D = \frac{2|A \bigcap B|}{|A| + |B|} \tag{9}$$

3.2.2. Hausdorff Distance

The Hausdorff Distance is mathematically defined as the maximum distance of a set to the nearest point in the other set, in other words, how close the segmentation and the expected output are. In most evaluations, we usually adopt the 95% Hausdorff Distance, Hausdorff95, which means the chosen distance is greater or equal to exactly 95% of the other distance in two point sets.

$$\overrightarrow{d_H}(A, B) = \max_{a \in A} \max_{b \in B} d(a, b) \tag{10}$$

$$H(A, B) = max\{\overrightarrow{d_H}(A, B), \overrightarrow{d_H}(B, A)\} \tag{11}$$

3.3. Experimental Results

To compare the model performance straightly, we gave the segmented mask generated by the baseline Naive-BoxSup method and our proposed method 3D-BoxSup, shown in **Figure 3**. Corresponding to the evaluated metric in section 3.2, the quantitative results are shown in **Table 1**. The chosen samples were randomly picked from HGG testing set and LGG testing set.

[1]https://www.med.upenn.edu/sbia/brats2018/data.html

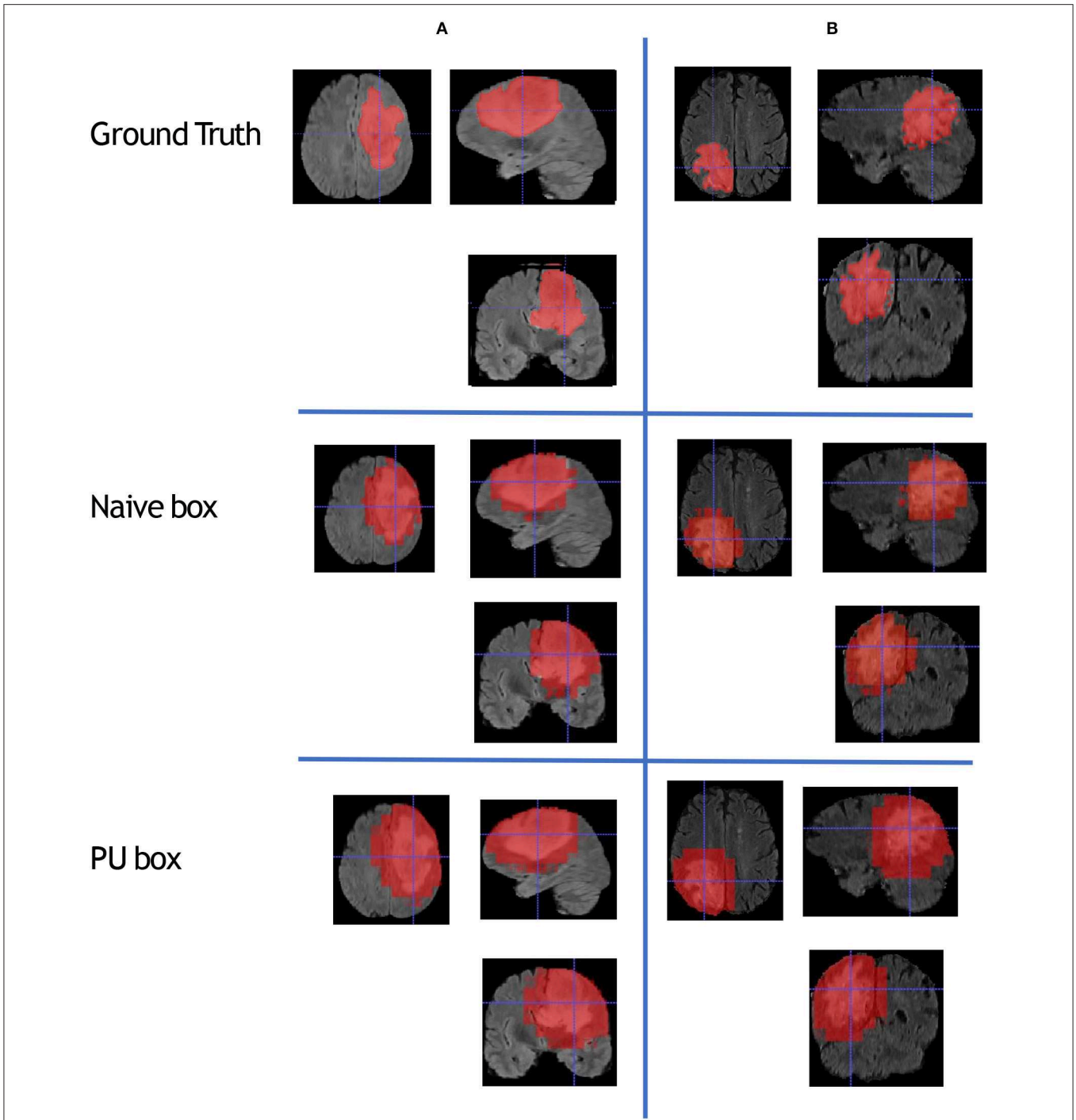


FIGURE 3 | This figure shows that we randomly sampled two patients from HGG testing set and LGG testing set respectively. Each column represents the applied method, and each row is the chosen patient. The A patient is from HGG samples, and the B patient is from LGG samples. The estimated segmentation result of WT is shown by both the naive method and our proposed PU box method. To better visualize the segmented result, we provide three different views: axial, coronal, and sagittal.

As can be seen from **Figure 3**, our proposed 3D-BoxSup method obviously produced a more accurate segmentation mask than the Naive-BoxSup method, which produced a much more noisy mask around the tumor boundary. It seems the Naive-BoxSup over-fits the data from no-tumor area S_0, which verifies that our method is able to alleviate this over-fitting and learning better from box area S_1.

In terms of the quantitative metrics of the Dice Score and Hausdorff Distance, our method performs better than the baseline method in each aspect, especially the Dice Score.

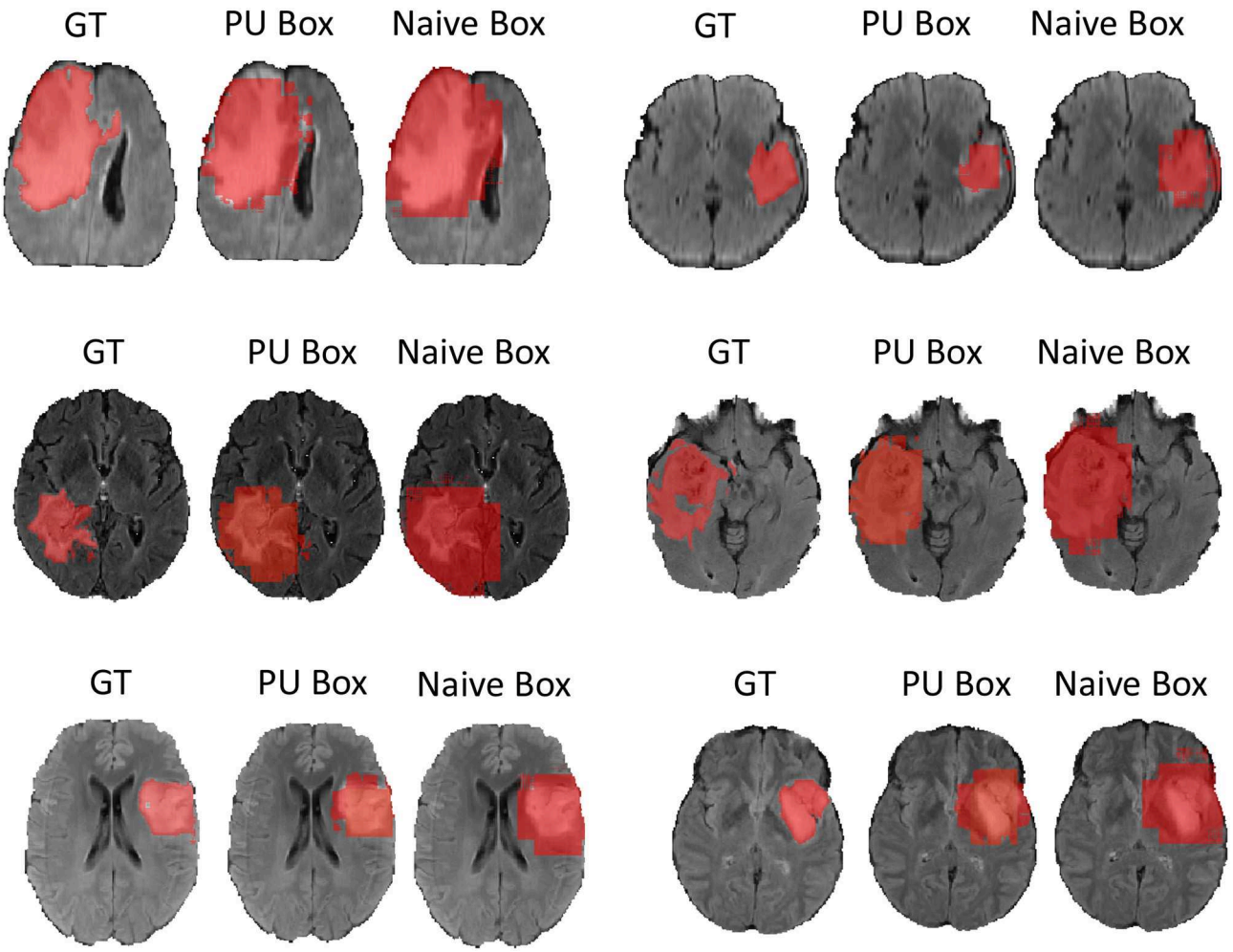


FIGURE 4 | We randomly choose six patients from testing set as more example displaying and for simplicity we only show the view of axial plane.

TABLE 1 | Mean values of Dice and Hausdorff measurements of the proposed method on the BraTS 2018 validation set.

	Dice	Hausdorff (mm)	Hausdorff95 (mm)
	WT	WT	WT
Naive-BoxSup (baseline)	0.49 ± 0.04	31.213 ± 2.316	20.857 ± 1.503
3D-BoxSup (ours)	0.62 ± 0.02	28.641 ± 1.395	15.476 ± 1.132
Region Grow	0.50	39.920	29.151

WT denotes whole tumor.

Visually, as shown in **Figure 4**, out method also generates finer segmentation mask than the baseline method. The variance of our 5-folds evaluation results is also smaller than the baseline model, which means our model is more robust. Also, due to the fact that we only applied a simple post-process for fill in the hole of segmented mask, the Hausdorff Distance could be influenced by the wrong segmentation area, which is beyond the tumor area. Compared with the hand-crafted region grow method (watershed clustering Ng et al., 2006), we set the threshold of discontinuities in gray-scale to be 0.5, and the result of the region grow is shown as below. To evaluate the region grow method, we tested it on both training data and testing data as tge region grow does not need to be trained. Overall, our proposed method shows a superiority when given a weak box annotation.

4. CONCLUSION

Precisely labeled data is limited in real world, especially for medical data; it would cost a significant amount time and labor to annotate the data, and it would also require a highly qualified doctor as the annotator. Thus, we need to refer to some easily labeled data, saving time, and also explore the information derived from these weakly labeled data. In this paper, we explored one of the possibilities of weakly supervised approach on medical image segmentation. Our method is called the "3D-BoxSup," which only acquired a 3D bounding box label for brain tumor.

Compared to the traditional supervised labeled data, which needs a fine boundary for tumor annotation, our annotated data is more accessible. However, training on the box labeled data would lead to over-fitting of the background as well as a biased risk function. Box labeled data is a typical positive-unlabeled task, and we thus proposed to apply the non-negative PU risk function (Kiryo et al., 2017) to boost the performance of our model. We have shown the effectiveness of our proposed method on the data provided by BRATS challenge (Menze et al., 2015). Since our model is a general method when tackling such box labeled data, our method can be further applied to mostly if not all of the segmentation tasks.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to BRATS challenge.

AUTHOR CONTRIBUTIONS

YX mainly implemented the method and conducted the experiments. MG provided the idea of this paper and contributed to the writing of the paper. JC and ZC substantially contributed to the revision. ZC performed the Region Growing experiments requested by the reviewers, and the JC helped with the writing of the revised paper. KB supervised the whole process, including the development of the concept, writing, revision, and other general advice.

FUNDING

This work was partially supported by NIH Award Number 1R01HL141813-01, NSF 1839332 Tripod+X, and SAP SE. We gratefully acknowledged the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. We were also grateful for the computational resources provided by Pittsburgh SuperComputing grant number TG-ASC170024.

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nature Scientific Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017c). Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. *The Cancer Imaging Archive.* doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *CoRR*, abs/1811.02629.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650.

Dai, J., He, K., and Sun, J. (2015). Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *CoRR*, abs/1503.01640.

Denis, F. (1998). "Pac learning from positive statistical queries," in *International Conference on Algorithmic Learning Theory* (Berlin; Heidelberg: Springer), 112–126.

Elkan, C., and Noto, K. (2008). "Learning classifiers from only positive and unlabeled data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, NV), 213–220. doi: 10.1145/1401890.1401920

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A. C., Bengio, Y., et al. (2015). Brain tumor segmentation with deep neural networks. *CoRR*, abs/1505.03540.

Isensee, F., Jaeger, P., Full, P. M., Wolf, I., Engelhardt, S., and Maier-Hein, K. H. (2017). Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. *CoRR*, abs/1707.00587.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. *CoRR*, abs/1802.10508.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S. G., Sinclair, M., Pawlowski, N., et al. (2017a). Ensembles of multiple models and architectures for robust brain tumour segmentation. *CoRR*, abs/1711.01468.

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017b). Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *CoRR*, abs/1412.6980.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. *CoRR*, abs/1703.00593.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans. Med. Imaging* 37, 2663–2674. doi: 10.1109/TMI.2018.2845918

Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Ng, H. P., Ong, S. H., Foong, K. W. C., Goh, P. S., and Nowinski, W. L. (2006). "Medical image segmentation using k-means clustering and improved watershed algorithm," in *Proceedings of the 2006 IEEE Southwest Symposium on Image Analysis and Interpretation*, SSIAI '06 (Washington, DC: IEEE Computer Society), 61–65.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in pytorch.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. *CoRR*, abs/1709.00382.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

TECHNOLOGY AND CODE
published: 29 April 2020
doi: 10.3389/fnins.2020.00125

BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice

Florian Kofler[1,2]*, *Christoph Berger*[1], *Diana Waldmannstetter*[1], *Jana Lipkova*[1], *Ivan Ezhov*[1], *Giles Tetteh*[1], *Jan Kirschke*[2], *Claus Zimmer*[2], *Benedikt Wiestler*[2†] and *Bjoern H. Menze*[1†]

[1] Image-Based Biomedical Modeling, Department of Informatics, Technical University of Munich, Munich, Germany, [2] Department of Neuroradiology, Klinikum rechts der Isar, Munich, Germany

OPEN ACCESS

Edited by:
Kamran Avanaki,
Wayne State University, United States

Reviewed by:
Suyash P. Awate,
Indian Institute of Technology Bombay, India
Guido Gerig,
NYU Tandon School of Engineering, United States

***Correspondence:**
Florian Kofler
florian.kofler@tum.de

[†] These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 30 September 2019
Accepted: 31 January 2020
Published: 29 April 2020

Citation:
Kofler F, Berger C, Waldmannstetter D, Lipkova J, Ezhov I, Tetteh G, Kirschke J, Zimmer C, Wiestler B and Menze BH (2020) BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. Front. Neurosci. 14:125. doi: 10.3389/fnins.2020.00125

Despite great advances in brain tumor segmentation and clear clinical need, translation of state-of-the-art computational methods into clinical routine and scientific practice remains a major challenge. Several factors impede successful implementations, including data standardization and preprocessing. However, these steps are pivotal for the deployment of state-of-the-art image segmentation algorithms. To overcome these issues, we present BraTS Toolkit. BraTS Toolkit is a holistic approach to brain tumor segmentation and consists of three components: First, the BraTS Preprocessor facilitates data standardization and preprocessing for researchers and clinicians alike. It covers the entire image analysis workflow prior to tumor segmentation, from image conversion and registration to brain extraction. Second, BraTS Segmentor enables orchestration of BraTS brain tumor segmentation algorithms for generation of fully-automated segmentations. Finally, Brats Fusionator can combine the resulting candidate segmentations into consensus segmentations using fusion methods such as majority voting and iterative SIMPLE fusion. The capabilities of our tools are illustrated with a practical example to enable easy translation to clinical and scientific practice.

Keywords: brain tumor segmentation, anonymization, MRI data preprocessing, medical imaging, brain extraction, BraTS, glioma

1. INTRODUCTION

Advances in deep learning have led to unprecedented opportunities for computer-aided image analysis. In image segmentation, the introduction of the U-Net architecture (Ronneberger et al., 2015) and subsequently developed variations like the V-Net (Milletari et al., 2016) or the 3D U-Net (Çiçek et al., 2016) have yielded algorithms for brain tumor segmentation that achieve a performance comparable to experienced human raters (Dvorak and Menze, 2015; Menze et al., 2015a; Bakas et al., 2018). A recent retrospective analysis of a large, multi-center cohort of glioblastoma patients convincingly demonstrated that objective assessment of tumor response via U-Net-based segmentation outperforms the assessment by human readers in terms of predicting patient survival (Kickingereder et al., 2019; Kofler et al., 2019), suggesting a potential benefit of implementing these algorithms into clinical routine.

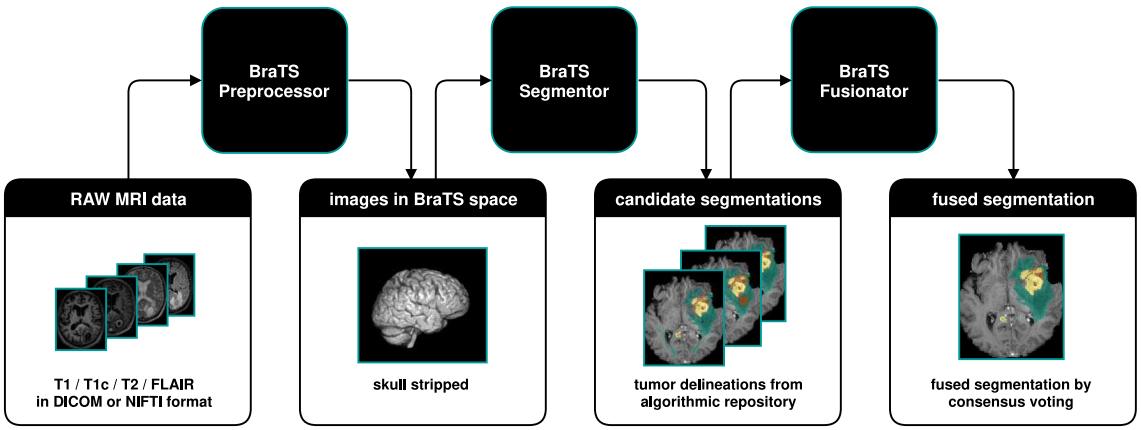


FIGURE 1 | Illustration of a typical dataflow to get from raw MRI scans to segmented brain tumors by combining the three components of the BraTS Toolkit. After preprocessing the raw MRI scans using the BraTS Preprocessor, the data is passed to the BraTS Segmentor, where arbitrary state-of-the-art models from the BraTS algorithmic repository can be used for segmentation. With BraTS Fusionator, multiple candidate segmentations may then be fused to obtain a consensus segmentation. As the Toolkit is designed to be completely modular and with clearly defined interfaces, each component can be replaced with custom solutions if required.

Recent works present diverse approaches toward brain tumor segmentation and analysis. Jena and Awate (2019) introduced a Deep-Neural-Network for image segmentation with uncertainty estimates based on Bayesian decision theory. Shboul et al. (2019) deployed feature-guided radiomics for glioblastoma segmentation and survival prediction. Jungo et al. (2018) analyzed the impact of inter-rater variability and fusion techniques for ground truth generation on uncertainty estimation. Shah et al. (2018) combined strong and weak supervision in training of their segmentation network to reduce overall supervision cost. Cheplygina et al. (2019) created an overview of Machine Learning methods in medical image analysis employing less or unconventional kinds of supervision.

In earlier years researchers experimented with a variety of approaches to tackle brain tumor segmentation (Prastawa et al., 2003; Menze et al., 2010, 2015b; Geremia et al., 2012), however in recent years the field is increasingly dominated by convolutional neural networks (CNN). This is also reflected in the contributions to the Multimodel Brain Tumor Segmentation Benchmark (BraTS) challenge (Bakas et al., 2018). The BraTS challenge (Menze et al., 2015a; Bakas et al., 2017) was introduced in 2012 at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), evaluating different algorithms for automated brain tumor segmentation. Therefore, every year the BraTS organizers provide a set of MRI scans, consisting of T1, T1c, T2, and FLAIR images from low- and high-grade glioma patients, coming with the corresponding ground truth segmentations.

Nonetheless, the computational methods presented in the BraTS challenge have not found their way into clinical and scientific practice. While the individual reasons vary, there are some key obstacles that impede the successful implementation of these algorithms. First of all, the availability of data for training, especially of high-quality, well-annotated data, is limited. Additionally, data protection as well as ethical barriers, complicate the development of centralized solutions, making local solutions strongly preferable. Furthermore, there are knowledge and skill barriers, when it comes to the conduction of setting up necessary preprocessing of data, while time and resources are limited.

While individual solutions for several of these problems exist, such as containerization for simplified distribution of code or public datasets, these are oftentimes fragmented and hence difficult to combine. Centralizing these efforts holds promise for making advances in image analysis easily available for broad implementation. Here we introduce three components to tackle these problems. First *BraTS preprocessor* facilitates data standardization and preprocessing for researchers and clinicians alike. Building upon that, varying tumor segmentations can be obtained from multiple algorithms with *BraTS Segmentor*. Finally, *BraTS Fusionator* can fuse these candidate segmentations into consensus segmentations by majority voting and iterative SIMPLE (Langerak et al., 2010) fusion. Together our tools represent *BraTS Toolkit* and enable a holistic approach integrating all the steps necessary for brain tumor image analysis.

2. METHODS

We developed BraTS Toolkit to get from raw DICOM data to fully automatically generate tumor segmentations in NIFTI format. The toolkit consists of three modular components. **Figure 1** visualizes how a typical brain tumor segmentation pipeline can be realized using the toolkit. The data is first preprocessed using the BraTS Preprocessor, then candidate segmentations are obtained from the BraTS Segmentor and finally fused via the BraTS Fusionator. Each component can be replaced with custom solutions to account for local

requirements[1]. A key design principle of the software is that all data processing happens locally to comply with data privacy and protection regulations.

BraTS Toolkit comes as a python package and can be deployed either via Python or by using the integrated command line interface (CLI). As the software is subject to ongoing development and improvement this work focuses on more abstract descriptions of the software's fundamental design principles. To ease deployment in scientific and clinical practice an up-to-date user guide with installation and usage instructions can be found here: https://neuronflow.github.io/BraTS-Toolkit/.

Users that prefer an easier approach can alternatively use the BraTS Preprocessor's graphical user interface (GUI) to take care of the data preprocessing[2]. The GUI is constantly improved in a close feedback loop with radiologists from the department of Neuroradiology at Klinikum Rechts der Isar (Technical University of Munich) to address the needs of clinical practitioners. Depending on the community's feedback, we plan to additionally provide graphical user interfaces for BraTS Segmentor and BraTS Fusionator in the future. Therefore, BraTS Toolkit features update mechanisms to ensure that users have access to the latest features.

2.1. Component One: BraTS Preprocessor

BraTS Preprocessor provides image conversion, registration, and anonymization functionality. The starting point to use BraTS Preprocessor is to have T1, T1c, T2, and FLAIR imaging data in NIFTI format. DICOM files can be converted to NIFTI format using the embedded dcm2niix conversion software (Li et al., 2016).

The main output of BraTS Preprocessor consists of the anonymized image data of all four modalities in BraTS space. Moreover, it generates the original input images converted to BraTS space, anonymized data in native space, defacing/skullstripping masks for anonymization, registration matrices to convert between BraTS and native space and overview images of the volumes' slices in png format. **Figure 2** depicts the data-processing in detail.

BraTS Preprocessor handles standardization and preprocessing of brain MRI data using a classical front- and back end software architecture. **Figure 3** illustrates the GUI variant's software architecture, which enables users without programming knowledge to handle MRI data pre-processing steps.

Advanced Normalization Tools (ANTs) (Avants et al., 2011) are deployed for linear registration and transformation of images into BraTS space, independent of the selected mode. In order to achieve proper anonymization of the image data there are four different processing modes to account for different local requirements and hardware configurations:

1. GPU brain-extraction mode
2. CPU brain-extraction mode
3. GPU defacing mode (under development)
4. CPU defacing mode

Brain extraction is implemented by means of HD-BET (Isensee et al., 2019) using GPU or CPU, respectively. HD-BET is a deep learning based brain extraction method, which is trained on glioma patients and therefore particularly well-suited for our task. In case the available RAM is not sufficient the CPU mode automatically falls back to ROBEX (Iglesias et al., 2011). ROBEX is another robust, but slightly less accurate, skull-stripping method that requires less RAM than HD-BET, when running on CPU.

Alternatively, the BraTS Preprocessor features GPU and CPU defacing modes for users who find brain-extraction too destructive. Defacing on the CPU is implemented via Freesurfer's mri-deface (Fischl, 2012), while deep-learning based defacing on the GPU is currently under development.

2.2. Component Two: BraTS Segmentor

The Segmentor module provides a standardized control interface for the BraTS algorithmic repository[3] (Bakas et al., 2018). This repository is a collection of Docker images, each containing a Deep Learning model and accompanying code designed for the BraTS challenge. Each model has a rigidly defined interface to hand data to the model and retrieve segmentation results from the model. This enables the application of state-of-the-art models for brain tumor segmentation on new data without the need to install additional software or to train a model from scratch. However, even though the algorithmic repository provides unified models, it is still up to the interested user to download and run each Docker image individually as well as manage the input and output. This final gap in the pipeline is closed by the Segmentor, which enables less experienced users to download, run and evaluate any model in the BraTS algorithmic repository. It provides a front end to manage all available containers and run them on arbitrary data, as long as the data conforms to the BraTS format. To this end, the Segmentor provides a command line interface to process data with any or all of the available Docker images in the repository while ensuring proper handling of the files. Its modular structure also allows anyone to extend the code, include other Docker containers or include it as a Python package.

2.3. Component Three: BraTS Fusionator

The Segmentor module can generate multiple segmentations for a given set of images which usually vary in accuracy and without prior knowledge, a user might be unsure which segmentation is the most accurate. The Fusionator module provides two methods to combine this arbitrary number of segmentation candidates into one final fusion which represents the consensus of all available segmentations. There are two main methods offered: Majority voting and the selective and iterative method for performance level estimation (SIMPLE) proposed by

[1]As an example users who do not want to generate tumor segmentations on their own hardware using the BraTS Segmentor, can alternatively try our experimental web technology based solution nicknamed the Kraken: https://neuronflow.github.io/kraken/.

[2]For an up-to-date installation and user guide please refer to: https://neuronflow.github.io/BraTS-Preprocessor/.

[3]https://github.com/BraTS/Instructions/blob/master/Repository_Links.md#brats-algorithmic-repository

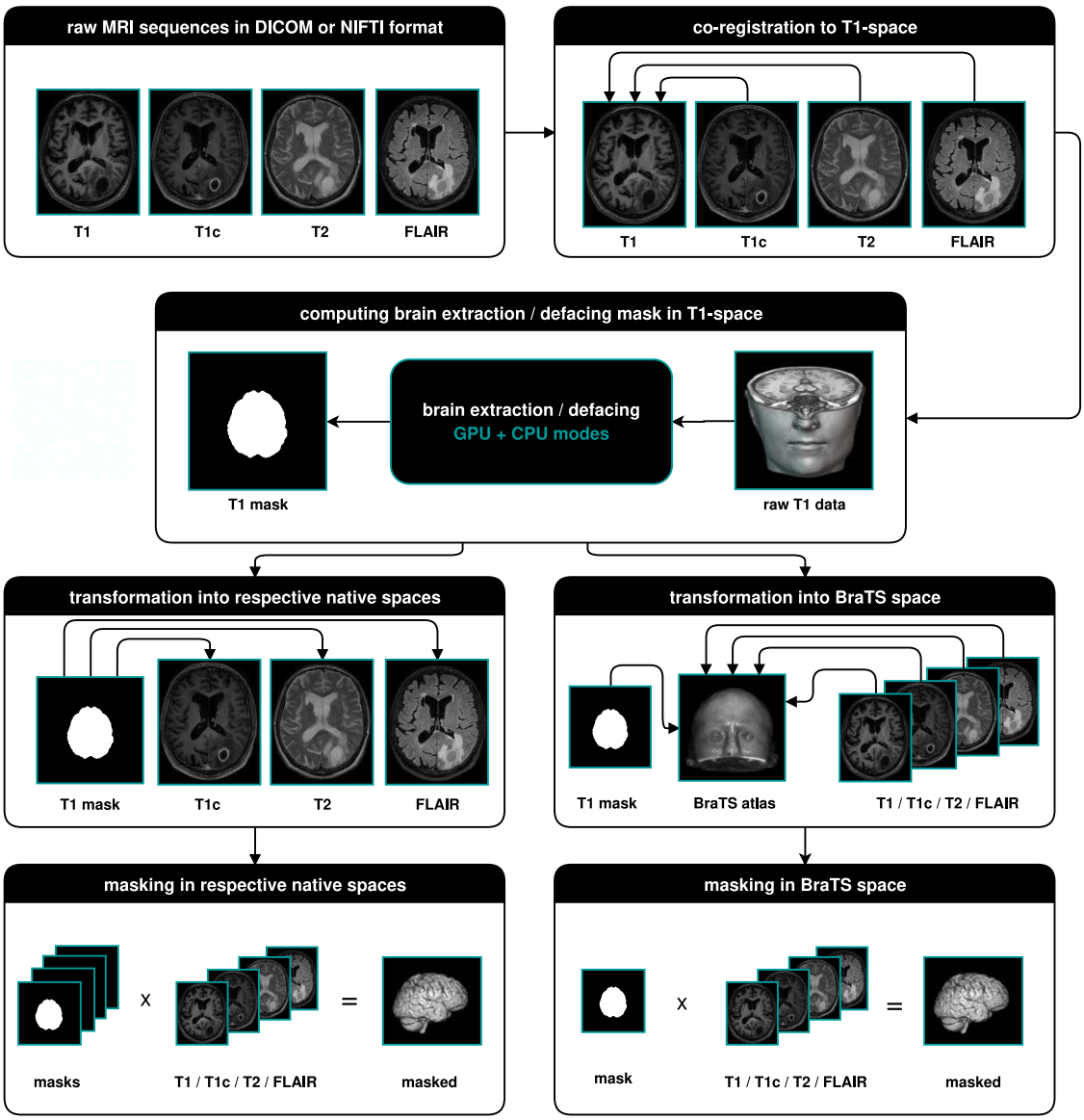


FIGURE 2 | Illustration of the data-processing. We start with a T1, T1c, T2, and FLAIR volume. In a first step we co-register all modalities to the T1 image. Depending on the chosen mode, we then compute the brain segmentation or defacing mask in T1-space. To morph the segmented images in native space, we transform the mask to the respective native spaces and multiply it with the volumes. For obtaining the segmented images in BraTS space, we transform the masks and volumes to the BraTS space using a brain atlas. We then apply the masks to the volumes.

Langerak et al. (2010). Both methods take all available candidate segmentations produced by the algorithms of the repository and combine each label to generate a final fusion. In majority voting, a class is assigned to a given voxel if at least half of the candidate segmentations agree that this voxel is of a certain class. This is repeated for each class to generate the complete segmentation. The SIMPLE fusion works as follows: First, a majority vote fusion with all candidate segmentations is performed. Secondly, each candidate segmentation is compared to the current consensus fusion and the resulting overlap score (a standard DICE measure in the proposed method) is used as a weight for the majority voting. This causes the candidate segmentations with higher estimated accuracy to have a higher influence on the final result. Lastly, each candidate segmentation with an accuracy below a certain threshold is dropped out after each iteration. This iterative process is stopped once the consensus fusion converges. After repeating the processes for each label, a final segmentation is obtained.

3. RESULTS

The broad availability of Python, Electron.js, and Docker allows us to support all major operating systems with an easy installation process. Users can choose to process data using the command line

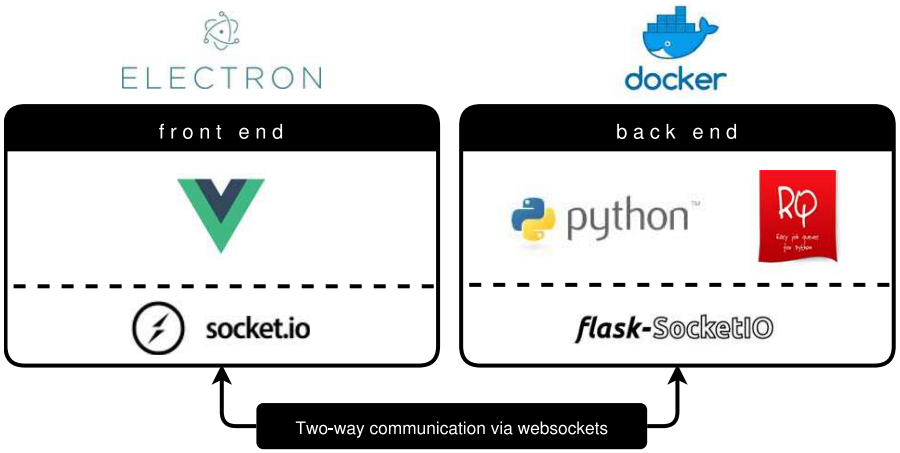


FIGURE 3 | BraTS Preprocessor software architecture (GUI variant). The front end is implemented by a Vue.js web application packaged via Electron.js. To ensure a constant runtime environment the Python based back end resides in a Docker container (Merkel, 2014). Redis Queue allows for load balancing and parallelization of the processing. The architecture enables two-way communication between front end and back end by implementing Socket.IO on the former and Flask-Socket.IO on the latter. In contrast to this the python package's front end is implemented using python-socketio.

(CLI) or through the user friendly graphical user interface (GUI). Depending on the available hardware, multiple threads are run to efficiently use the system's resources.

3.1. Practicality in Clinical and Scientific Practice

To test the practicality of BraTS Toolkit we conducted a brain tumor segmentation experiment on 191 patients of the BraTS 2016 dataset. As a first step we generated candidate tumor segmentations. BraTS Segmentor allowed us to rapidly obtain tumor delineations from ten different algorithms of the BraTS algorithmic repository (Bakas et al., 2018). The standardized user interface of BraTS Segmentor abstracts all the required background knowledge regarding docker and the particularities of the algorithms. In the next step we used BraTS Fusionator to fuse the generated segmentations by consensus voting. **Figure 4** shows that fusion by iterative SIMPLE and class-wise majority voting had a slight advantage over single algorithms. This effect was particularly driven by removal of false positives as illustrated for an exemplary patient in **Figure 5**. BraTS Toolkit enabled us to conduct the experiment in a user-friendly way. With only a few lines of Python code we were able to obtain segmentation results in a fully-automated fashion. This impression was confirmed by experiments on further in house data-sets where we also deployed the CLI and GUI variants of all three BraTS Toolkit components with great feedback from clinical and scientific practitioners. Users especially appreciated the increased robustness and precision of consensus segmentations compared to existing single algorithm solutions.

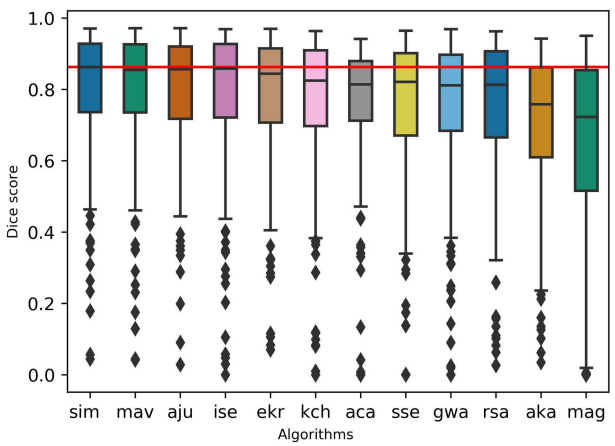


FIGURE 4 | Evaluation of the segmentation results on the BraTS 2016 data set for whole tumor labels on $n = 191$ evaluated test cases. We generated candidate segmentations with ten different algorithms. Segmentation methods are sorted in descending order by mean dice score. The two fusion methods, iterative SIMPLE (sim) and class-wise majority voting displayed on the left, outperformed individual algorithms depicted further right. The red horizontal line shows the SIMPLE median dice score ($M = 0.863$) for better comparison.

4. DISCUSSION

Overall, the BraTS Toolkit is a step toward the democratization of automatic brain tumor segmentation. By lowering resource and knowledge barriers, users can effectively disseminate dockerized brain tumor segmentation algorithms collected through the BraTS challenge. Thus, it makes objective brain tumor volumetry, which has been demonstrated to be superior to traditional image assessment (Kickingereder et al., 2019), readily available for scientific and clinical use.

Currently, BraTS segmentation algorithms and therefore BraTS Segmentor require each of T1, T1c, T2, and FLAIR sequences to be present. In practice, this can become a limiting factor due to errors in data acquisition or incomplete protocols leading to missing modalities. Recent efforts try to bridge this

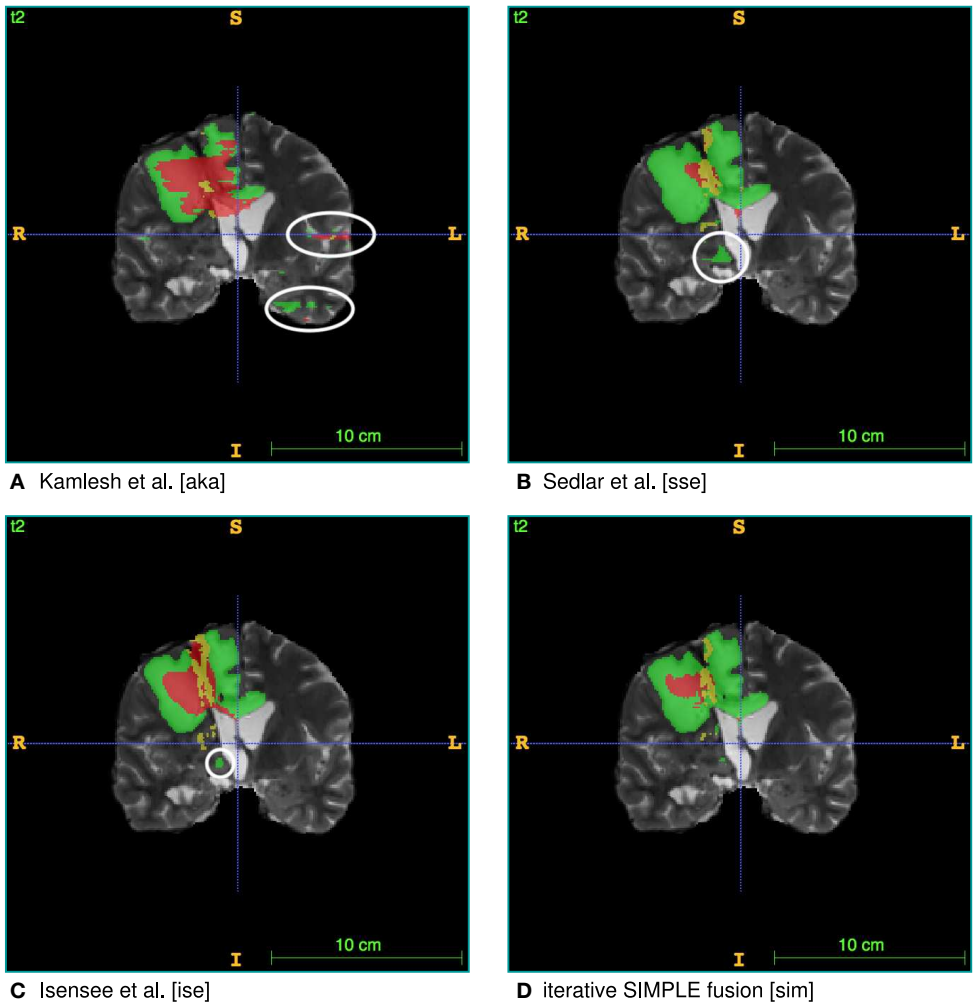


FIGURE 5 | Single algorithm vs. iterative SIMPLE consensus segmentation. T2 scans with segmented labels by exemplary candidate algorithms from **(A)** Pawar et al. (2018), **(B)** Sedlar (2018), and **(C)** Isensee et al. (2017) (Green: edema; Red: necrotic region/non-enhancing tumor; Yellow: enhancing tumor). **(D)** Shows a consensus segmentation obtained using the iterative SIMPLE fusion. Notice the false positives marked with white circles on the candidate segmentations. These outliers are effectively reduced in the fusion segmentation shown in **(D)**.

gap by using machine learning techniques to reconstruct missing image modalities (e.g., Dorent et al., 2019; Li et al., 2019).

Other crucial aspects of data preprocessing are the lack of standards for pulse sequences across different scanners and manufacturers, and absence of data acquisition protocols' harmonization in general. For the moment, we address this only with primitive image standardization strategies as described in **Figure 2**. However, in clinical and scientific practice, we already found our application to be very robust across different data sources. Brain extraction with HD-BET also proved to be sound for patients from multiple institutions with different pathologies (Isensee et al., 2019).

These limitations are in fact some of the key motivations for our initiative. We strive to provide researchers with tools to build comprehensive databases which capture more of the data variability in magnetic resonance imaging. In the longterm this will enable the development of more precise algorithms. With BraTS Toolkit clinicians can actively contribute to this process.

Through well-defined interfaces, the resulting output from our software can be integrated seamlessly with further downstream software to create new scientific and medical applications such as but not limited to, fully-automatic MR reporting[4] or tumor growth modeling (Ezhov et al., 2019; Lipková et al., 2019). Another promising future direction is to focus on integration with the local PACS to enable streamlined processing of imaging data directly from the radiologist's workplace.

[4]Our Kraken web service can be seen as an an exemplary prototype for this (for the moment it is not for clinical use, but for research and entertainment purposes only). The Kraken is able to send automatically generated segmentation and volumetry reports to the user's email address: https://neuronflow.github.io/kraken/.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

FK conceptualized the BraTS Toolkit, programmed the BraTS Preprocessor and contributed to paper writing. CB programmed and conceptualized the BraTS Fusionator and BraTS Segmentor and contributed to paper writing. DW, JL, IE, and JK conceptualized the BraTS Preprocessor and contributed to paper writing. GT conceptualized the BraTS Preprocessor software architecture and contributed to paper writing. CZ conceptualized the BraTS Preprocessor and provided feedback on the BraTS Fusionator. BW and BM conceptualized the BraTS Toolkit and contributed to paper writing.

FUNDING

BM, BW, and FK are supported through the SFB 824, subproject B12. Supported by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. With the support of the Technical University of Munich–Institute for Advanced Study, funded by the German Excellence Initiative.

REFERENCES

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *NeuroImage* 54, 2033–2044. doi: 10.1016/j.neuroimage.2010.09.025

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). 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*.

Cheplygina, V., de Bruijne, M., and Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Analysis* 54, 280–296. doi: 10.1016/j.media.2019.03.009

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 424–432.

Dorent, R., Joutard, S., Modat, M., Ourselin, S., and Vercauteren, T. (2019). "Hetero-modal variational encoder-decoder for joint modality completion and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019* eds D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan (Cham: Springer International Publishing), 74–82.

Dvorak, P., and Menze, B. (2015). "Local structure prediction with convolutional neural networks for multimodal brain tumor segmentation", in *Medical Computer Vision: Algorithms for Big Data*, (Cham: Springer International Publishing), 59–71.

Ezhov, I., Lipkova, J., Shit, S., Kofler, F., Collomb, N., Lemasson, B., et al. (2019). "Neural parameters estimation for brain tumor growth modeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer International Publishing), 787–795.

Fischl, B. (2012). Freesurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Geremia, E., Menze, B. H., Ayache, N. (2012). "Spatial decision forests for glioma segmentation in multi-channel mr images," in *MICCAI Challenge on Multimodal Brain Tumor Segmentation*, (Citeseer), 34.

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. Med. Imaging*, 30, 1617–1634. doi: 10.1109/TMI.2011.2138152

Isensee, F., Kickingereder, P., Bonekamp, D., Bendszus, M., Wick, W., Schlemmer, H.-P., et al. (2017). "Brain tumor segmentation using large receptive field deep convolutional neural networks," in *Bildverarbeitung für die Medizin 2017* (Springer), 86–91.

Isensee, F., Schell, M., Tursunova, I., Brugnara, G., Bonekamp, D., Neuberger, U., et al. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, (Wiley Online Library), 40, 4952–964. doi: 10.1002/hbm.24750

Jena, R., and Awate, S. P. (2019). "A bayesian neural net to segment images with uncertainty estimates and good calibration," in *International Conference on Information Processing in Medical Imaging* (Springer), 3–15.

Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., et al. (2018). "On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer International Publishing), 682–690.

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., et al. (2019). Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 20, 728–740. doi: 10.1016/S1470-2045(19)30098-1

Kofler, F., Paetzold, J., Ezhov, I., Shit, S., Krahulec, D., Kirschke, J., et al. (2019). "A baseline for predicting glioblastoma patient survival time with classical statistical models and primitive features ignoring image information," in *International MICCAI Brainlesion Workshop* (Springer).

Langerak, T. R., van der Heide, U. A., Kotte, A. N., Viergever, M. A., Van Vulpen, M., Pluim, J. P., et al. (2010). Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *IEEE Trans. Med. Imaging* 29, 2000–2008. doi: 10.1109/TMI.2010.2057442

Li, H., Paetzold, J. C., Sekuboyina, A., Kofler, F., Zhang, J., Kirschke, J. S., et al. (2019). "Diamondgan: unified multi-modal generative adversarial networks for mri sequences synthesis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, eds D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan (Cham: Springer International Publishing), 795–803.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* 264, 47–56. doi: 10.1016/j.jneumeth.2016.03.001

Lipková, J., Angelikopoulos, P., Wu, S., Alberts, E., Wiestler, B., Diehl, C., et al. (2019). Personalized radiotherapy design for glioblastoma: integrating mathematical tumor models, multimodal scans, and bayesian inference. *IEEE Trans. Med. Imaging* 38, 1875–1884. doi: 10.1109/TMI.2019.2902044

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015a). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Menze, B. H., Van Leemput, K., Lashkari, D., Riklin-Raviv, T., Geremia, E., Alberts, E., et al. (2015b). A generative probabilistic model and discriminative extensions for brain lesion segmentation—with application to tumor and stroke. *IEEE Trans. Med. Imaging* 35, 933–946. doi: 10.1109/TMI.2015.2502596

Menze, B. H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., and Golland, P. (2010). "A generative model for brain tumor segmentation in multi-modal images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Berlin; Heidelberg: Springer Berlin Heidelberg), 151–159.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux J.* 2014:2.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (IEEE), 565–571.

Pawar, K., Chen, Z., Shah, N. J., and Egan, G. (2018). "Residual encoder and convolutional decoder neural network for glioma segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 263–273.

Prastawa, M., Bullitt, E., Moon, N., Van Leemput, K., and Gerig, G. (2003). Automatic brain tumor segmentation by subject specific modification of atlas priors1. *Acad. Radiol.* 10, 1341–1348. doi: 10.1016/S1076-6332(03)00506-3

Ronneberger, O., Fischer, P., Brox, and Thomas. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Sedlar, S. (2018). "Brain tumor segmentation using a multi-path cnn based method," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer International Publishing), 403–422.

Shah, M. P., Merchant, S., and Awate, S. P. (2018). "Ms-net: mixed-supervision fully-convolutional networks for full-resolution segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 379–387.

Shboul, Z. A., Alam, M., Vidyaratne, L., Pei, L., Elbakary, M. I., and Iftekharuddin, K. M. (2019). Feature-guided deep radiomics for glioblastoma patient survival prediction. *Front. Neurosci.* 13:966. doi: 10.3389/fnins.2019.00966

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 04 August 2020
doi: 10.3389/fncom.2020.00061

Overall Survival Prediction in Glioblastoma With Radiomic Features Using Machine Learning

Ujjwal Baid[1†], Swapnil U. Rane[2†], Sanjay Talbar[1], Sudeep Gupta[3], Meenakshi H. Thakur[4], Aliasgar Moiyadi[5] and Abhishek Mahajan[4]*

[1] Department of Electronics and Telecommunication Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India, [2] Department of Pathology, Tata Memorial Centre, ACTREC, HBNI, Navi-Mumbai, India, [3] Department of Medical Oncology, Tata Memorial Centre, ACTREC, HBNI, Navi-Mumbai, India, [4] Department of Radiodiagnosis and Imaging, Tata Memorial Centre, Tata Memorial Hospital, HBNI, Mumbai, India, [5] Department of Neurosurgery Services, Tata Memorial Centre, Tata Memorial Hospital, HBNI, Mumbai, India

Glioblastoma is a WHO grade IV brain tumor, which leads to poor overall survival (OS) of patients. For precise surgical and treatment planning, OS prediction of glioblastoma (GBM) patients is highly desired by clinicians and oncologists. Radiomic research attempts at predicting disease prognosis, thus providing beneficial information for personalized treatment from a variety of imaging features extracted from multiple MR images. In this study, first-order, intensity-based volume and shape-based and textural radiomic features are extracted from fluid-attenuated inversion recovery (FLAIR) and T1ce MRI data. The region of interest is further decomposed with stationary wavelet transform with low-pass and high-pass filtering. Further, radiomic features are extracted on these decomposed images, which helped in acquiring the directional information. The efficiency of the proposed algorithm is evaluated on Brain Tumor Segmentation (BraTS) challenge training, validation, and test datasets. The proposed approach achieved 0.695, 0.571, and 0.558 on BraTS training, validation, and test datasets. The proposed approach secured the third position in BraTS 2018 challenge for the OS prediction task.

Keywords: brain tumor, glioblastoma, overall survival, radiomic, machine learning

OPEN ACCESS

Edited by:
Spyridon Bakas, University of Pennsylvania, United States

Reviewed by:
Madhura Ingalhalikar, Symbiosis International University, India
Ahmad Chaddad, Guilin University of Electronic Technology, China
Benedikt Wiestler, Technical University of Munich, Germany

***Correspondence:**
Abhishek Mahajan
drabhishek.mahajan@yahoo.in

[†] These authors have contributed equally to this work

Received: *20 November 2019*
Accepted: *27 May 2020*
Published: *04 August 2020*

Citation:
Baid U, Rane SU, Talbar S, Gupta S, Thakur MH, Moiyadi A and Mahajan A (2020) Overall Survival Prediction in Glioblastoma With Radiomic Features Using Machine Learning. Front. Comput. Neurosci. 14:61. doi: 10.3389/fncom.2020.00061

INTRODUCTION

Glioblastoma (GBM) remains the most aggressive primary malignant brain tumor in adults, with a median survival time of 15 months and 5-year survival of ∼5% after initial diagnosis (Chang et al., 2016). Nearly all patients with GBM relapse despite providing maximal safe surgical resection, radiotherapy, temozolomide, and aggressive therapy. Spatial and temporal intra-tumor heterogeneity, extent, and location are some of the factors that make these tumors challenging to resect and, in some cases, inoperable. The inability to perform complete surgical tumor resection and poor drug delivery to the brain contributes notably to the lack of effective treatment and poor prognosis (Mahajan et al., 2015).

Certain biological variables such as MGMT promoter methylation status, 1p/19q deletion, and IDH1 gene mutation status have been shown to explain to a certain extent this observed variation, in addition to certain host variables such as age and gender. The fact that GBM shows extremely wide clinical behavior points to the fact that the current understanding of GBM as a single disease entity is an oversimplification. This is further supported by the fact that there have been multiple

attempts to divide GBM into more distinct subgroups using molecular subtyping (Verhaak et al., 2010). However, these methods are difficult to replicate in routine clinical practice owing to the complexity of the assays and high costs. Further, tumors show subtype plasticity with a complex transition from one subtype to another during progression (Lee et al., 2018). Thus, predicting survival of patients with GBM is a challenging task.

Magnetic resonance imaging (MRI) plays a vital role in neuro-oncology for initial diagnosis and assessment of treatment response and is increasingly used as a powerful non-invasive predictive tool. Researchers have identified that MRI provides distinct information that can predict survival independently of pathologic and clinical data. The process that extracts various quantitative features on the basis of intensity, volume, shape, and textural variations from radiographic images and design predictive algorithms to find the association of these vast features to the survival and outcome of the patient is known as radiomics (Chaddad et al., 2019b). Radiomics incorporates several essential disciplines, including radiology for imaging interpretation, computer vision for quantitative feature extraction, and machine learning for classifier evaluation and regression (Seow et al., 2018; Vaidya et al., 2019).

In recent years, several radiomic models have been proposed for survival prediction (Huang et al., 2016), distant metastasis prediction (Coroller et al., 2015), and molecular characteristics classification (Kickingereder et al., 2016a,b). Researchers extracted several radiomic features on the basis of texture, area, volume, and Euler characteristics-based features from different intra-tumor parts (Shboul et al., 2019). Extreme Gradient Boosting (XGBoost) was used as a regressor to predict the OS. This approach achieved 0.519 accuracy on Brain Tumor Segmentation (BraTS) 2018 test dataset. In another study, multi-planer spatial convolutional neural networks were used for brain tumor segmentation, and semantic and agnostic features were extracted on these segmented tumor parts. These radiomic features were provided as input to multilayer perceptron (MLP) to predict OS (Banerjee et al., 2019). Although the proposed approach performed well for segmentation task, the algorithm performed poorly on BraTS 2018 test dataset for overall survival (OS) prediction task. Other than the sophisticated machine learning approaches, a simple linear regressor was used on only nine features. These features were computed by the volume, by summing up the voxels and the surface area, and by summing up the magnitude of the gradients along with three directions. There were fewer chances of overfitting because of only nine features, and hence, the method performed well (Feng et al., 2019). Multi-scale texture features-based approach for predicting GBM patients' progression-free survival and OS on T1- and T2-weighted fluid-attenuated inversion recovery (FLAIR) MRIs was proposed using the random forest (Chaddad et al., 2018). The study results showed that the identified seven-feature set, when combined with clinical factors, improved the model performance, yielding an area under the receiver operating characteristic curve (AUC) value of 85.54% for OS predictions. Osman et al. extracted a set of 147 radiomic image features locally from three tumor subregions on standardized preoperative multiparametric MR images. LASSO regression was applied for identifying an informative subset of chosen features, whereas a Cox model was used to obtain the coefficients of those selected features (Osman, 2019). Despite the various correlations between imaging features, genomic expression, and survival reported in the literature, no single analysis has been substantive enough to enter clinical practice.

In another study, usefulness of geometric shape features, extracted from MR images, as a potential non-invasive way to characterize GBM tumors and predict the OS times of patients with GBM, is evaluated (Chaddad et al., 2016a). Multi-contrast MRI texture features were used for the prediction of survival of patients GBM using texture features derived from gray-level co-occurrence matrices (GLCMs). The statistical analysis based on the Kaplan–Meier method and log-rank test was conducted in order to identify the texture features most closely associated with the OS (Chaddad and Tanougast, 2016; Chaddad et al., 2016b). A study underlines that radiomic features could be complimentary to biopsy-based sequencing methods to predict survival of patients with IDH1 wild-type GBM (Chaddad et al., 2019a).

This study aims to evaluate the efficiency of the radiomic feature-based MRI signatures from multi-modal MRI data and to find their associations with OS in patients with high-grade gliomas (HGGs) with improved accuracy compared with those of the available state-of-the-art methods. The rest of the manuscript is organized as follows: the dataset used for the study, preprocessing steps, and radiomic feature extraction framework is described in the *Material and Method*. Sample results are discussed in the *Result and Discussion*. *Conclusion and Future Work* concludes the paper with future direction.

MATERIALS AND METHODS

We participated in BraTS 2018 challenge, which mainly focused on two tasks:

1. segmentation of brain tumor with intra-tumor parts like edema, enhancing tumor, and necrotic part; and
2. OS prediction of the patients in days with the help of imaging features.

In this study, we mainly focused on the survival prediction aspects of the BraTS challenge.

Dataset

Since 2012, every year, the BraTS challenge is organized at Medical Image Computing and Computer Assisted Intervention (MICCAI) conference (Menze et al., 2015). The challenge is to segment HGG and low-grade glioma (LGG) with high accuracy. From 2017 onwards, the task is extended for the prediction of OS of the patients in days as well (Bakas et al., 2017a,b, 2019). The BraTS organizers had provided multi-institutional training dataset of 163 patients diagnosed with GBM. For the validation dataset and test dataset, 53 and 130 cases were provided separately. The data were obtained from various institutions all over the globe, with different clinical protocols and scanners. For each patient, MRI data of size 240 × 240 × 155 were provided with FLAIR, T1, T1ce, and T2

modalities along with the ground truth as shown in **Figure 1**. The same annotation protocol was followed to segment all the cases manually by one to four raters, which were later verified by expert neurologists with more than 15 years of experience. The labels were termed as edema, enhancing tumor (ET), and necrosis. One of the tasks of BraTS 2018 challenge was to auto-segment the tumor into its three constituent regions, namely,

1. enhancing tumor region (ET), which shows hyperintensity in T1 postcontrast when compared with T1;
2. tumor core (TC), which entails the ET, necrotic (fluid filled), and non-enhancing (solid) parts; and
3. whole tumor (WT), which includes all intra-tumor parts along with edema.

Additional information like resection status, age, and survival in days were also provided exclusively for OS prediction task. The

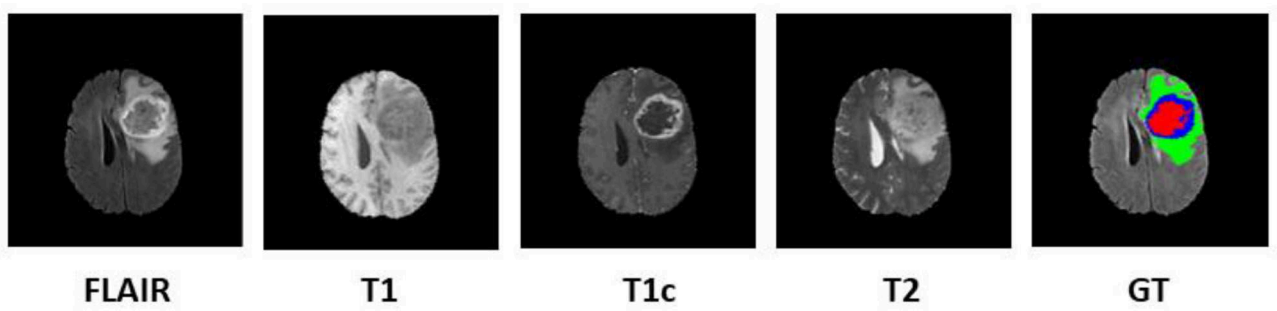


FIGURE 1 | Multi-modal data with four channels provided in BraTS 2018 challenge dataset along with ground truth (GT). Subtumor parts are represented as follows: green, edema; blue, enhancing tumor; red, necrosis. BraTS, Brain Tumor Segmentation; FLAIR, fluid-attenuated inversion recovery.

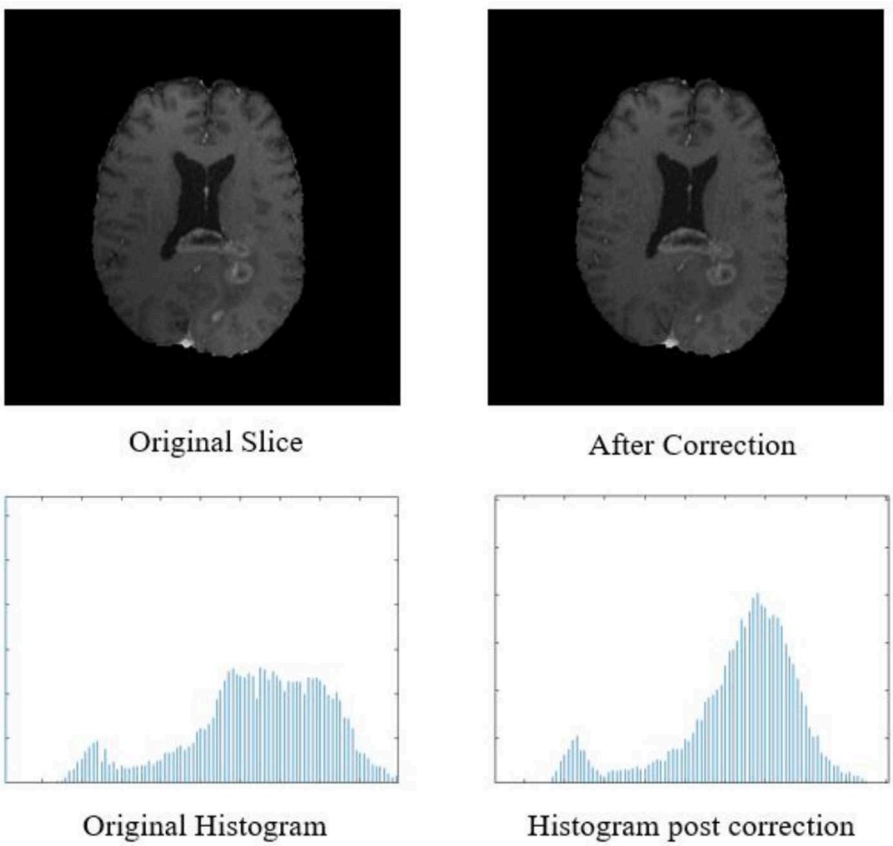


FIGURE 2 | Top row: original input MR slice and slice after biased field correction. Bottom row: corresponding histograms of original slice and histogram after biased field correction. The horizontal *X*-axis of the histogram is intensity, and the vertical *Y* axis is frequency.

MR data provided by BraTS organizers was skull stripped and co-registered to 1 mm × 1 mm × 1 mm isotropic resolution. The proposed three-step pipeline is shown in **Figure 2**.

Proposed Methodology

We proposed three-step methodology for OS prediction, as shown in **Figure 3**. In our approach, radiomic features were extracted on region of interest (ROI). The segmentation labels were provided for training dataset only, and hence, we segmented the tumors in validation and test dataset first and then extracted the radiomic features on the segmented ROI as first step. In step 2, radiomic features were extracted, and feature selection and OS prediction model was designed in step 3.

Preprocessing

The biased field algorithm was applied on FLAIR, T1, T2, and T1ce channel to correct the intensity inhomogeneity with N4ITK tool (Tustison et al., 2010). From **Figure 2**, it can be observed that the slice after bias field correction is more homogeneous in terms of intensity. All the four MR channels were normalized to zero mean and unit variance. We extracted multi-channel and multi-regional radiomic features with the help of intra-tumor annotations provided by the organizers on training dataset. We segmented the tumor with patch-based 3D U-Net architecture (Baid et al., 2019, 2020).

Radiomic Feature Extraction

We extracted radiomic features on FLAIR and T1ce channels. Because whole tumor is best seen in FLAIR modality and enhancing tumor boundaries can be best visualized in T1ce modality, we selected these channels only for feature extraction. We computed radiomic features on these modalities with three varying combinations of intra-tumor parts as a whole tumor, that is, all intra-tumor parts, necrosis with enhancing tumor, and

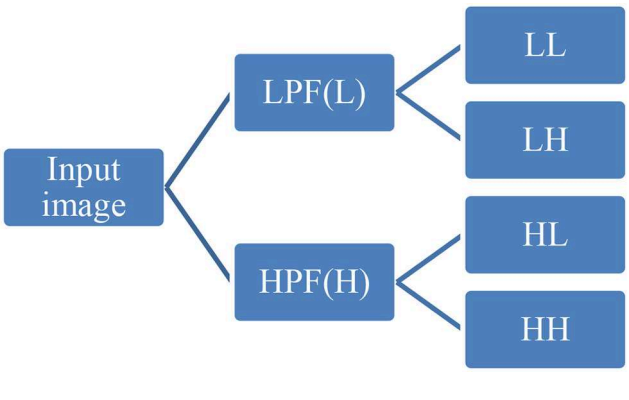


FIGURE 4 | Representative diagram for stationary wavelet decomposition. LPF, low-pass filter; HPF, high-pass filter.

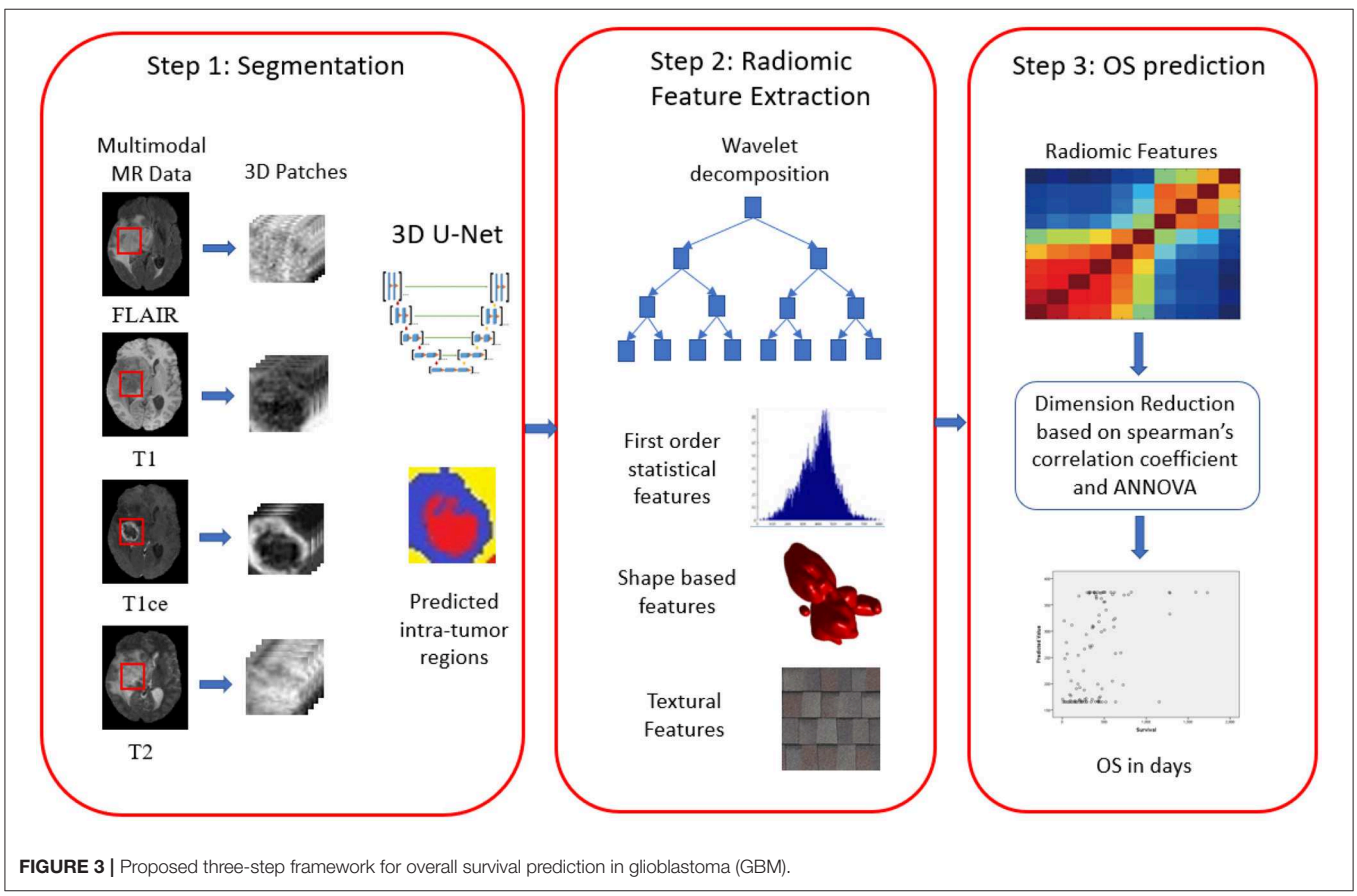


FIGURE 3 | Proposed three-step framework for overall survival prediction in glioblastoma (GBM).

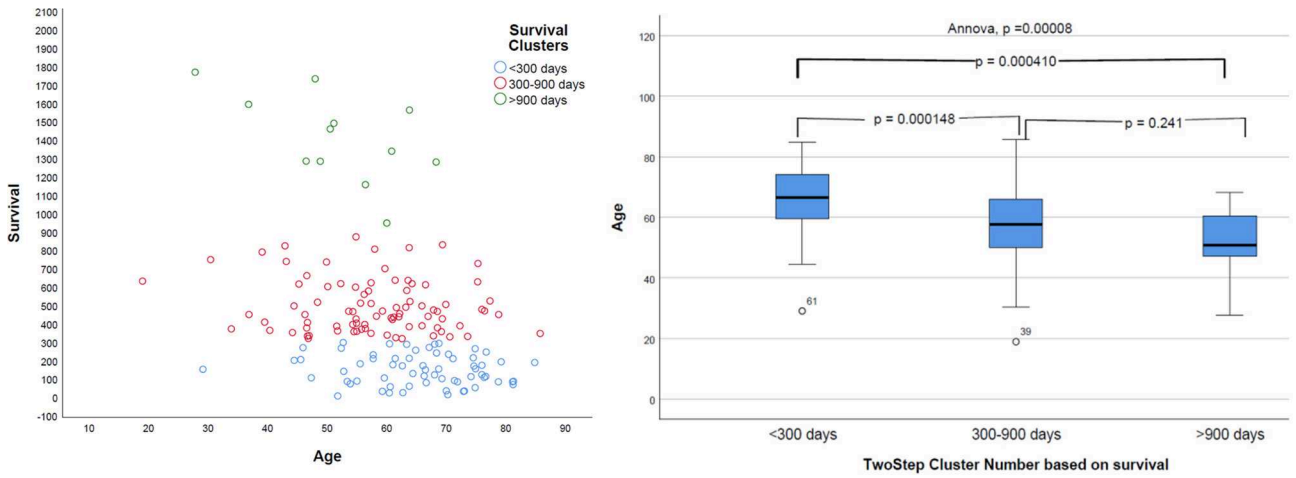


FIGURE 5 | Distribution of survival groups identified using two-step clustering and correlation with age. Left: the age is in years and overall survival in given days. Right: on the *X*-axis of two-step cluster, Group 1, Group 2, and Group 3 are represented.

enhancing tumor only. So at the end, we had six combinations to extract radiomic features.

First-order statistical features and shape-based features were extracted from these combinations of ROI and MR channels. ROI was decomposed into four sub-bands with a multi-level 2-D stationary wavelet decomposition using a biorthogonal wavelet (Kickingereder et al., 2016b). In the first step, ROI was decomposed into two sub-bands with low-pass filter (LPF) and high-pass filter (HPF). Further, these sub-bands were again passed through LPF and HPF, giving LL, LH, HL, and HH bands (**Figure 4**). This was to extract directional texture features from approximate, horizontal, vertical, and diagonal components obtained after decomposition of the ROI (Nason and Silverman, 1995). It should be noted that we have not down-sampled the LPF and HPF to generate LL, LH, HL, and HH purposefully to avoid any sort of loss of information. GLCM features were extracted from these sub-bands (Haralick and Shanmugam, 1973). One hundred and thirteen first-order statistics, shape-based, and GLCM features were extracted for each tumor part and modality considering all four wavelet sub-bands. Thus, we had a total of 678 radiomic features extracted from six different combinations of tumor parts and modalities. Each patient in the BraTS dataset was provided with age as additional information, which we had concatenated in our feature vector. Finally, for each patient, we had 679 variables to be used to train the regression model for the survival prediction task. In training dataset, we had 163 patients for whom OS was provided in days. The radiomic feature extraction pipeline is available at Github[1].

Survival Prediction

Survival prediction was divided into two tasks. One task aimed at classifying patients into three survival groups obtained by unsupervised two-step clustering. These groups roughly correspond to the known survival groups in GBM (PMID: 22517216). The survival groups were characterized as long survivors (e.g., >900 days), short survivors (e.g., <300 days), and mid-survivors (e.g., between 300 and 900 days). For precise treatment planning, it is valuable to categorize a patient to either of these survival subgroups. This will enable clinicians to decide how aggressively a patient needs to be treated. The second task aimed at predicting OS in days, which is the same as the task required for BraTS 2018.

Delineating Survival Groups

Natural grouping of patients based on survival was investigated using unsupervised two-step hierarchical clustering. This resulted in three groups with a good silhouette of separation: Group 1 (short survivors; patients with <300 days' OS, $n = 65$), Group 2 (mid-survivors; patients with OS between 300 and 900 days, $n = 86$), and Group 3 (long survivors; patients with >900 days' survival, $n = 12$). Pearson's correlation revealed a strong inverse relationship of age with this group, with younger patients having the greatest OS ($p = 0.000008$, $r = -0.368$), as shown in **Figure 5**.

Assessment of Relationship With Survival and Radiomic Feature Vector Dimension Reduction

In order to reduce the dimensionality of the feature vector, Spearman's correlation coefficient was calculated for each pair of radiomic features. The features having Spearman's correlation coefficient >0.95 with each other were discarded, retaining a single feature in each set (**Supplementary Table 1**). This reduced the feature vector size from 679 to 118. The feature set was further reduced to 54 by excluding all variables with statistically insignificant ($p > 0.05$) relationship with the survival groups (tested using ANOVA) identified above and with OS (tested using Pearson's correlation coefficient). It was observed that in terms of normalized importance,

[1]https://github.com/ujjwalbaid0408/Radiomics

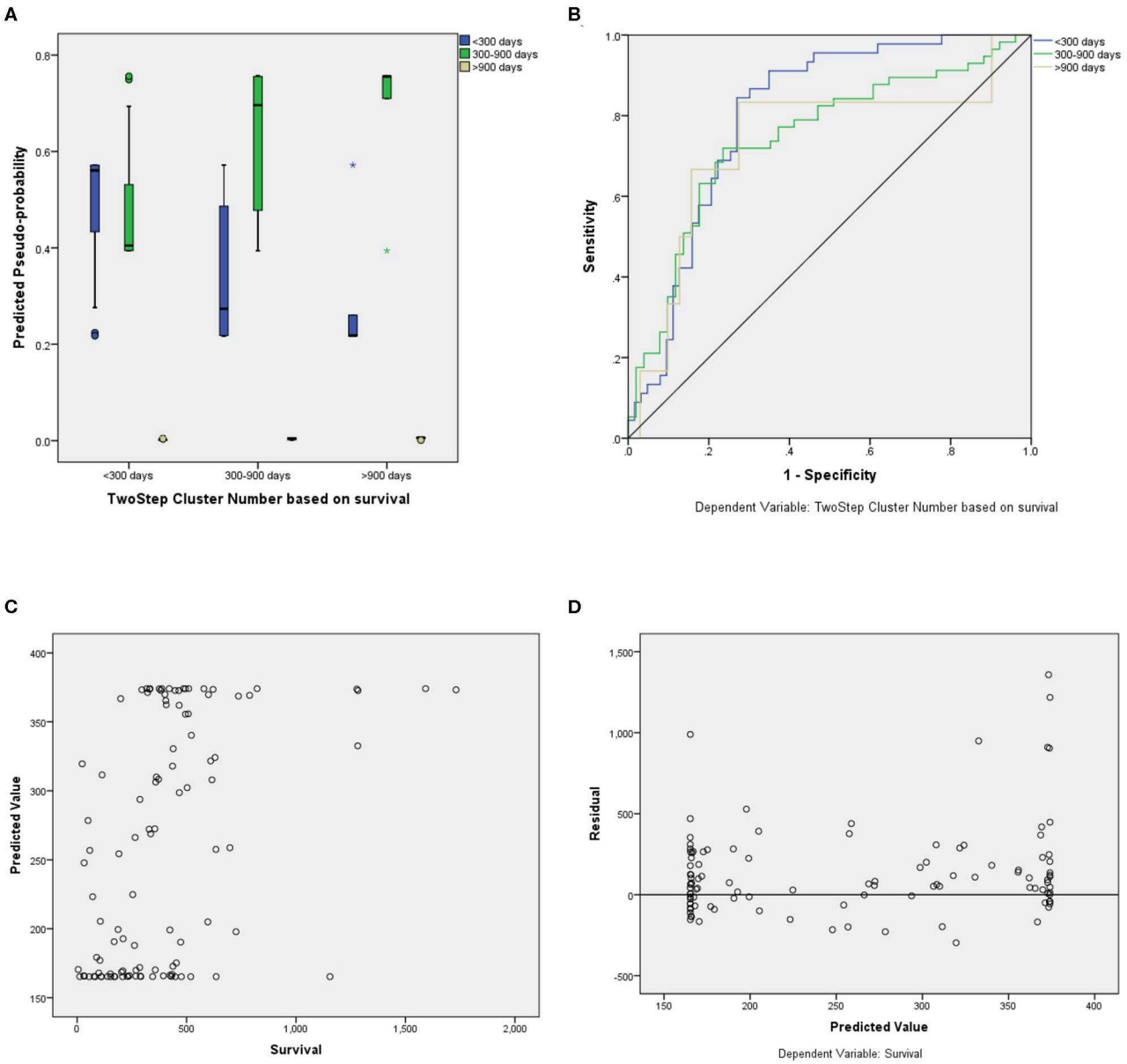


FIGURE 6 | Performance summary of the overall survival (OS) prediction algorithm. **(A)** Predicted pseudoprobabilities across the three prediction categories. **(B)** Area under the receiver operating characteristic curve (AUC) for the three categories. **(C)** Dot plot of predicted and actual survival in days. **(D)** Residual vs. predicted plots for the survival prediction in days.

age is the most important feature. Because whole tumor is visible in FLAIR modality with hyperintense pixels, their features followed age. The enhancement tumor and core tumor counts were of significant importance for survival prediction (**Supplementary Table 1**).

Predicting Survival Groups Using Radiomic Features

Neural networks were designed using MLP to build a predictive model using the reduced radiomic feature vector set and age.

Neural Network Design

A single neural network was designed to classify the features in the three survival categories and to predict the OS in days. The neural network designed had two hidden layers. The number of units per layer were fixed to "auto." The sigmoid activation function was used in hidden layers and output layers. Results were replicated by setting a random seed. For fair evaluation and to avoid overfitting, the BraTS training dataset was further divided into training (51.5%), validation (14.7%), and testing (33.7%) subsets by using randomly generated Bernoulli variates.

TABLE 1 | Quantitative evaluation of multilayer perceptron for OS prediction on the BraTS dataset.

BraTS dataset	Accuracy	MSE	Median SE	Std. deviation	Spearman *R*
Training	0.695	18,920.841	9,139.551	22,253.812	0.877
Validation	0.571	59,550,213.1	1,136,111.6	128,250,465.8	0.427
Testing	0.558	338219.366	38408.16	939986.796	0.222

OS, overall survival; BraTS, Brain Tumor Segmentation; MSE, mean squared error.

TABLE 2 | Quantitative evaluation of MLP and RF for OS prediction on BraTS validation dataset.

Approach	Accuracy	MSE	Median SE	Std. deviation	Spearman *R*
RF	0.375	6,109,105.6	47,545.13	143,070.37	0.11
MLP	0.571	5,955,021.1	11,361.6	12,825,046.8	0.427

MLP, multilayer perceptron; RF, random forest; OS, overall survival; BraTS, Brain Tumor Segmentation; MSE, mean squared error.

All the features were rescaled with an adjusted normalized correction of 0.2. We also performed an individual variable importance analysis (**Supplementary Table 1**).

RESULT AND DISCUSSION

Neural Network Performance

For the prediction of survival categories, the neural network demonstrated an accuracy of 70.2% in the training subset and 62.5 and 63.6% in the validation and testing subsets, respectively, which we divided from BraTS training dataset. The accuracy was 73% for the entire training dataset. The AUC was 0.799 (0.817 for Group 1, 0.709 for Group 2, and 0.784 for Group 3). A summary of the model performance is shown in **Figure 6**. The designed model performed better for patients in the mid-survivor groups, with the least accuracy for patients in the long-survivor group.

For fair evaluation of all the proposed algorithms of researchers participating in the BraTS challenge, organizers had provided an online evaluation platform. Participants were expected to submit the results on this platform, and later, they could download the quantitative results for the same[2]. It had been observed that despite less accuracy on validation dataset, our method achieved the third position in OS prediction task in the BraTS challenge[3]. The most convincing reason behind this was that all other participants might have overfitted their methods to the validation dataset. BraTS organizers have provided leaderboard of all the participants with segmentation and OS prediction task with several quantitative evaluation matrices[4]. We evaluated the proposed approach on BraTS training testing and validation dataset as shown in **Table 1**. The comparison between random forest and MLP is given in **Table 2** on BraTS validation dataset.

We have also evaluated the efficiency of the proposed approach with 10-fold validation. At every fold, 90% of patients are used for training and 10% of samples are kept as a holdout. The classification accuracy at each fold is given in **Figure 7**. The *X*-axis of the plot represents the fold number, and *Y*-axis gives the corresponding classification accuracy. The average accuracy is found to be 58.49, which is comparable with accuracy on BraTS 2018 validation dataset, which proves the robustness of the proposed approach.

DISCUSSION

Predicting outcomes has been the holy grail of modern oncology, notoriously difficult to achieve with high accuracy, yet driving numerous investigators toward finding newer ways of attempting to reach that goal. Because of multiple challenging factors, this task is out of clinical reach. Some of the challenges are the limitations of the human mind and recording devices to quantify the biological variations. The other is an amalgamation of cross-disciplinary interactions of clinical sides (for treatment planning) and engineering sides (toward quantitative analysis).

In this work, we have evaluated a couple of methods (MLP and RF) to predict OS using radiomic feature extracted using deep learning-based segmentation and feature pipeline as shown in **Figure 3**. What is interesting is the fact that although our segmentation pipeline did not feature in the top models submitted to BraTS 2018, our survival prediction pipeline did make it to the third position on the basis of the performance metrics decided by the BraTS 2018 organizers. The neural network achieved an accuracy of 0.583 with a relatively low standard error. Age was the most important variable in the predictive model. Further, we have also identified radiomic features that contributed maximally to the model.

The importance of the independent variable in descending order is given in **Supplementary Table 1**. It was observed that age is the most important factor for the OS prediction. We observed that radiomic analysis of tumor core region, which is comprised of necrosis and enhancing tumor on FLAIR modality contributed significantly toward prediction OS. It can be concluded that the core tumor count, that is, volume of enhancing tumor and tumor core, are of extreme importance in OS prediction of patients with GBM.

The amount of clinical information (only age and OS) provided in BraTS 2018 is extremely limited; and no details regarding gender, other co-existing comorbidities, performance status, and details of treatment received are provided. Considering these limiting factors, it is interesting to note that radiomic features coupled with age could explain a significant amount of variability seen in the OS of these patients with GBM. Although predicting survival in terms of closest number of days is desirable, in actual clinical practice, it often suffices to predict prognostically relevant groups for treatment intensification. For example, we were able to identify patients with $<$300 days of survival with a significantly high accuracy (0.804). These patients are ideal candidates for treatment intensification. Our accuracy in predicting survival for the long survivors was the least, possibly owing to the small number of cases in that group.

[2]https://ipp.cbica.upenn.edu/
[3]https://www.med.upenn.edu/sbia/brats2018/rankings.html
[4]https://www.cbica.upenn.edu/BraTS18/

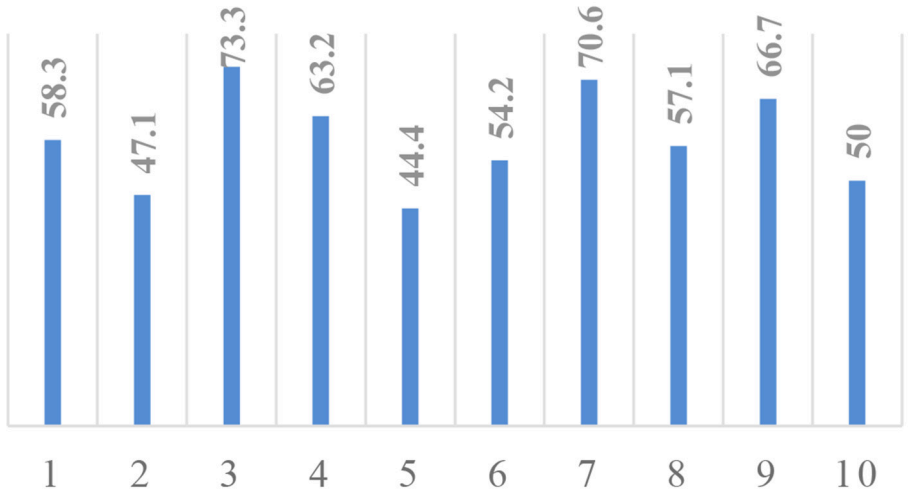


FIGURE 7 | *K*-fold cross-validation analysis. *X*-axis, fold number; *Y*-axis, accuracy.

Implementation Details

The radiomic feature extraction pipeline is designed in MATLAB environment. The neural network design, feature reduction, and other statistical analysis were performed with SPSS v24 on computing machine with Windows 10 operating system.

CONCLUSION AND FUTURE WORK

In this study, we evaluated the efficiency of radiomic features and machine learning-based classifier to predict the OS of the patients diagnosed with GBM. Multi-modal radiomic features were extracted from the FLAIR and T1ce channel of preoperated MRI data. OS of the patient was predicted with MLP and RF regressors. The classification accuracy shows that MLP outperformed over random forest in terms of accuracy. The proposed approach achieved the third position in BraTS 2018. We have also identified radiomic features that contribute maximally to the neural network's predictive ability. Further, the work could potentially include incorporating additional prognostic variables such as pathologic assessment information, molecular aberration information, comorbidities, and performance status into the predictive model.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The datasets analyzed for this study can be found in the BraTS 2018 dataset https://www.med.upenn.edu/sbia/brats2018/data.html.

AUTHOR CONTRIBUTIONS

UB, AM, and SR conducted the experiment. All authors contributed to writing the manuscript and are responsible.

FUNDING

This study was funded by the Ministry of Electronics and Information Technology, India through the Visvesvaraya Ph.D. Scheme of Electronics and Information Technology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00061/full#supplementary-material

REFERENCES

Baid, U., Talbar, S., Rane, S., Gupta, S., Thakur, M. H., Moiyad, A., et al. (2019). "Deep learning radiomics algorithm for gliomas (DRAG) model: a novel approach using 3D UNET based deep convolutional neural network for predicting survival in gliomas," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Granada), 369–379.

Baid, U., Talbar, S., Rane, S., Gupta, S., Thakur, M. H., Moiyadi, A., et al. (2020). A novel approach for fully automatic intra-tumor segmentation with 3D U-Net architecture for gliomas. *Front Comput Neurosci.* 14:10. doi: 10.3389/fncom.2020.00010

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4, 1–13. doi: 10.1038/sdata.2017.117

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Arch.* 286. doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Reyes, M., Jakab, A., Bauer, S., and Rempfler, M. (2019). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv Prepr. arXiv*1811.01328.

Banerjee, S., Mitra, S., and Shankar, B. U. (2019). "Multi-planar spatial-ConvNet for segmentation and survival prediction in brain cancer." in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Granada), 94–104.

Chaddad, A., Daniel, P., Sabri, S., Desrosiers, C., and Abdulkarim, B. (2019a). Integration of radiomic and multi-omic analyses predicts survival of newly diagnosed IDH1 wild-type glioblastoma. *Cancers.* 11, 1–16. doi: 10.3390/cancers11081148

Chaddad, A., Desrosiers, C., Hassan, L., and Tanougast, C. (2016a). A quantitative study of shape descriptors from glioblastoma multiforme phenotypes for predicting survival outcome. *Br. J. Radiol.* 89:20160575. doi: 10.1259/bjr.20160575

Chaddad, A., Desrosiers, C., and Toews, M. (2016b). Radiomic analysis of multi-contrast brain MRI for the prediction of survival in patients with glioblastoma multiforme. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2016:4035–4038. doi: 10.1109/EMBC.2016.7591612

Chaddad, A., Kucharczyk, M. J., Daniel, P., Sabri, S., Jean-Claude, B. J., Niazi, T., et al. (2019b). Radiomics in glioblastoma: current status and challenges facing clinical implementation. *Front. Oncol.* 9:374. doi: 10.3389/fonc.2019.00374

Chaddad, A., Sabri, S., Niazi, T., and Abdulkarim, B. (2018). Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. *Med. Biol. Eng. Comput.* 56, 2287–2300. doi: 10.1007/s11517-018-1858-4

Chaddad, A., and Tanougast, C. (2016). Extracted magnetic resonance texture features discriminate between phenotypes and are associated with overall survival in glioblastoma multiforme patients. *Med. Biol. Eng. Comput.* 54, 1707–1718. doi: 10.1007/s11517-016-1461-5

Chang, K., Zhang, B., Guo, X., Zong, M., Rahman, R., Sanchez, D., et al. (2016). Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. *Neuro. Oncol.* 18, 1680–1687. doi: 10.1093/neuonc/now086

Coroller, T. P., Grossmann, P., Hou, Y., Rios Velazquez, E., Leijenaar, R. T., Hermann, G., et al. (2015). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* 114, 345–350. doi: 10.1016/j.radonc.2015.02.015

Feng, X., Tustison, N. J., Patel, S. H., and Meyer, C. H. (2019). "Brain tumor segmentation using an ensemble of 3D U-nets and overall survival prediction using radiomic features," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Granada), 279–288.

Haralick, R. M., and Shanmugam, K. (1973). Textural features of image classification. *IEEE Trans. Syst. Man Cybern.* 3, 610–621.

Huang, Y., Liu, Z., He, L., Chen, X., Pan, D., Ma, Z., et al. (2016). Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. *Radiology* 281, 947–957. doi: 10.1148/radiol.2016152234

Kickingereder, P., Bonekamp, D., Nowosielski, M., Kratz, A., Sill, M., Burth, S., et al. (2016a). Radiogenomics of glioblastoma: machine learning–based classification of molecular characteristics by using multiparametric and multiregional mr imaging features. *Radiology* 281, 907–918. doi: 10.1148/radiol.2016161382

Kickingereder, P., Burth, S., Wick, A., Götz, M., Eidel, O., Schlemmer, H. P., et al. (2016b). Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology* 280, 880–889. doi: 10.1148/radiol.2016160845

Lee, E., Yong, R. L., Paddison, P., and Zhu, J. (2018). Comparison of glioblastoma (GBM) molecular classification methods. *Semin. Cancer Biol.* 53, 201–211. doi: 10.1016/j.semcancer.2018.07.006

Mahajan, A., Moiyadi, V. A., Jalali, R., and Sridhar, E. (2015). Radiogenomics of glioblastoma: a window into its imaging and molecular variability. *Cancer Imaging* 15(Suppl. 1):P14. doi: 10.1186/1470-7330-15-S1-P14

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging.* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Nason, G. P., and Silverman, B. W. (1995). "The stationary wavelet transform and some statistical applications." in *Wavelets and Statistics*, eds A. Antoniadis and G. Oppenheim (New York, NY: Springer New York), 281–299.

Osman, A. F. I. (2019). A multi-parametric MRI-based radiomics signature and a practical ML model for stratifying glioblastoma patients based on survival toward precision oncology. *Front. Comput. Neurosci.* 13:58. doi: 10.3389/fncom.2019.00058

Seow, P., Wong, J. H. D., Ahmad-Annuar, A., Mahajan, A., Abdullah, N. A., and Ramli, N. (2018). Quantitative magnetic resonance imaging and radiogenomic biomarkers for glioma characterisation: a systematic review. *Br. J. Radiol.* 91:20170930. doi: 10.1259/bjr.20170930

Shboul, Z. A., Alam, M., Vidyaratne, L., Pei, L., and Iftekharuddin, K. M. (2019). "Glioblastoma survival prediction," in *Brainlesion Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Quebec City, QC), 508–515.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Vaidya, T., Agrawal, A., Mahajan, S., Thakur, M. H., and Mahajan, A. (2019). The continuing evolution of molecular functional imaging in clinical oncology: the road to precision medicine and radiogenomics (Part I). *Mol. Diagnosis Ther.* 23, 27–51. doi: 10.1007/s40291-018-0367-3

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 14 December 2020
doi: 10.3389/fncom.2020.563439

Can Tumor Location on Pre-treatment MRI Predict Likelihood of Pseudo-Progression vs. Tumor Recurrence in Glioblastoma?—A Feasibility Study

Marwa Ismail[1], Virginia Hill[2,3], Volodymyr Statsevych[2], Evan Mason[2], Ramon Correa[1], Prateek Prasanna[4], Gagandeep Singh[1], Kaustav Bera[1,5], Rajat Thawani[1], Manmeet Ahluwalia[6], Anant Madabhushi[1,7] and Pallavi Tiwari[1]*

[1] Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, United States, [2] Department of Neuroradiology, Imaging Institute, Cleveland Clinic, Cleveland, OH, United States, [3] Department of Radiology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, [4] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, United States, [5] Maimonides Medical Center, New York, NY, United States, [6] Brain Tumor and Neuro-Oncology Center, Cleveland Clinic, Cleveland, OH, United States, [7] Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, OH, United States

OPEN ACCESS

Edited by:
Bjoern Menze, Technical University of Munich, Germany

Reviewed by:
Benedikt Wiestler, Technical University of Munich, Germany
Anahita Fathi Kazerooni, University of Pennsylvania, United States
Hassan Mohy-ud-Din, Lahore University of Management Sciences, Pakistan

***Correspondence:**
Marwa Ismail
mxi125@case.edu

Received: *18 May 2020*
Accepted: *19 November 2020*
Published: *14 December 2020*

Citation:
Ismail M, Hill V, Statsevych V, Mason E, Correa R, Prasanna P, Singh G, Bera K, Thawani R, Ahluwalia M, Madabhushi A and Tiwari P (2020) Can Tumor Location on Pre-treatment MRI Predict Likelihood of Pseudo-Progression vs. Tumor Recurrence in Glioblastoma?—A Feasibility Study. Front. Comput. Neurosci. 14:563439. doi: 10.3389/fncom.2020.563439

A significant challenge in Glioblastoma (GBM) management is identifying pseudo-progression (PsP), a benign radiation-induced effect, from tumor recurrence, on routine imaging following conventional treatment. Previous studies have linked tumor lobar presence and laterality to GBM outcomes, suggesting that disease etiology and progression in GBM may be impacted by tumor location. Hence, in this feasibility study, we seek to investigate the following question: Can tumor location on treatment-naïve MRI provide early cues regarding likelihood of a patient developing pseudo-progression vs. tumor recurrence? In this study, 74 pre-treatment Glioblastoma MRI scans with PsP (33) and tumor recurrence (41) were analyzed. First, enhancing lesion on Gd-T_{1w} MRI and peri-lesional hyperintensities on T_{2w}/FLAIR were segmented by experts and then registered to a brain atlas. Using patients from the two phenotypes, we construct two atlases by quantifying frequency of occurrence of enhancing lesion and peri-lesion hyperintensities, by averaging voxel intensities across the population. Analysis of differential involvement was then performed to compute voxel-wise significant differences (p-value < 0.05) across the atlases. Statistically significant clusters were finally mapped to a structural atlas to provide anatomic localization of their location. Our results demonstrate that patients with tumor recurrence showed prominence of their initial tumor in the parietal lobe, while patients with PsP showed a multi-focal distribution of the initial tumor in the frontal and temporal lobes, insula, and putamen. These preliminary results suggest that lateralization of pre-treatment lesions toward certain anatomical areas of the brain may allow to provide early cues regarding assessing likelihood of occurrence of pseudo-progression from tumor recurrence on MRI scans.

Keywords: glioblastoma, tumor recurrence, pseudo-progression, atlas, ADIFFI

INTRODUCTION

A significant challenge in management of Glioblastoma (GBM), the most aggressive form of brain cancer, is differentiating tumor recurrences from pseudo-progression (PsP) on routine magnetic resonance (MR) scans (Parvez et al., 2014). PsP is a benign radiation-induced treatment effect which occurs in approximately 19–33% of all malignant brain tumors (Wang et al., 2016) and usually stabilizes or regresses without further treatment. Unfortunately, PsP mimics tumor recurrence radiologically on routine MRI scans [Gadolinium-enhanced T1-weighted (Gd-T1w), T2-weighted (T2w), FLAIR], making it challenging to differentiate from true tumor recurrence (Wang et al., 2016). Studies have previously explored advanced imaging modalities such as perfusion imaging (Prager et al., 2015; Chuang et al., 2016; Detsky et al., 2017), MR spectroscopy (Chuang et al., 2016), and diffusion-weighted imaging (Prager et al., 2015) in distinguishing tumor recurrence from PsP. However, these advanced imaging modalities are limited by acquisition variability, costs, reproducibility, and unavailability at most clinical sites (Brandsma et al., 2008). Reliable disease assessment using routine imaging is thus needed in order to aid in accurately identifying PsP from tumor recurrence. Timely identification of these conditions could avoid unnecessary interventions in patients with PsP, while allowing for change in treatment for patients with tumor recurrence (Parvez et al., 2014).

Multiple studies have linked initial lesion location in the brain to be a prognostic marker of tumor recurrence and overall survival in diffuse Gliomas (Ellingson et al., 2012). For instance, recent studies have demonstrated a higher rate of 1p19q deletion in the frontal lobe (Laigle-Donadey et al., 2004), and absence of 1DH1 mutation within the insula (Metellus et al., 2010). Similarly, Gliomas in the frontal locations have been shown to be associated with a better prognosis compared to other locations (Stockhammer et al., 2012). Further, enhancing lesion developing in the periventricular region has been linked to PsP (Patel et al., 2014; Van West et al., 2017). These studies seem to suggest that the underlying disease etiology may be driven by tumor location. Hence, it may be reasonable to rationalize that initial GBM location in the brain may implicitly contribute to an increased likelihood of a patient developing pseudo-progression or tumor recurrence, following conventional treatment of maximal surgical resection and chemo-radiation therapy.

In this feasibility study, we evaluate this hypothesis that lesion location on pre-treatment MR scans could provide early cues regarding likelihood of a patient developing tumor recurrences vs. PsP. In order to anatomically localize the disease, we employ "population atlases" of GBM phenotypes to establish predisposition of tumor recurrence or PsP to specific spatial locations in the brain based on their frequency of occurrence (Larjavaara et al., 2007; Ellingson et al., 2012; Bilello et al., 2016). The statistical population atlases allow for the succinct encapsulation of structural and anatomical variability of the disease across a patient population using a single reference or canonical representation. We will construct population atlases on a cohort of 74 brain MRI scans across two lesion sub-compartments (peritumoral hyperintensities as defined on FLAIR scans and enhancing core as defined on T_{1w} MRI), to quantify the frequency of occurrence of PsP and tumor recurrence in pre-treatment lesions. We will further employ a statistical mapping technique, ADIFFI, to identify if there exist any statistically significant lesion locations in the brain across the two disease pathologies, by comparing the population atlases of PsP and tumor recurrence.

MATERIALS AND METHODS

Study Population

The Institutional Review Board-approved and HIPAA-compliant study comprised GBM patient population from Cleveland Clinic. The population cohort for pre-treatment cases included 74 cases in total; 41 tumor recurrence cases, and 33 PsP cases. All cases were confirmed for disease presence using the criteria provided below. Informed consent was obtained for all patients involved in the study. All MR scans were acquired using either a 1.5 Tesla or a 3-Tesla scanner. **Table 1** summarizes the demographics for this study population.

Confirmation for Disease Presence

Our dataset was identified by performing a retrospective review of all brain tumor patients who had an enhancing lesion within 6 months of treatment (treatment strategies for each patient are provided in **Supplementary Material**). Our inclusion criteria consisted of the following: (1) pre-, and post-treatment MRI scans that are of diagnostic image quality as determined by collaborating radiologists, (2) availability of all 3 routine MRI sequences (Gd-T1w, T2w, FLAIR), (3) a suspected post-treatment enhancing lesion with more than 5 millimeters (mm) of rim or nodular enhancement, and (4) confirmation of PsP or tumor recurrence for the suspected lesion.

In order to carefully assess the presence of PsP/recurrence, the following steps were followed. First, MRI and other advanced imaging scans (if available) were read by a neuro-radiologist (board-certified in neuroradiology, CAQ) to identify the presence of PsP/recurrence using image assessment based on the RANO criteria (Wen et al., 2010). Then, the initial interpretation was reviewed by patient's clinical team (Neuro-oncology staff, radiation oncologist). All cases were later discussed at a multidisciplinary tumor board in order to provide the final decision. The tumor board constituted of 2+ neuro-oncologists, a neuro-radiologist, a neuropathologist, and one or more surgeons, at our collaborating institution (CCF). A consensus opinion on each individual case was finally formed based on a methodical review of the clinical assessment, prior therapies, and assessment based on imaging features, to identify every study as PsP or tumor recurrence.

Image Registration and Tumor Segmentation

Manual segmentations in our work were carefully performed by our collaborating experts, where every 2-D slice of each MRI scan (for $n = 74$ studies) with visible tumor was manually annotated [using 3D Slicer (Kikinis et al., 2014)] into 2 regions: enhancing lesion and T_{2w}/FLAIR hyperintense peri-lesional component.

TABLE 1 | Summary of the study population used in this work to create population atlases for PsP and tumor recurrence.

Characteristic	Tumor recurrence	Pseudo-progression
No. of patients	41	33
Females	16	12
Males	25	21
Mean age (year)	59.1	61.96
Age range (year)	26–75	24–75

Gd-T_{1w} MRI scans were used to delineate the enhancing lesion, while both T_{2w} and FLAIR scans were used to annotate the T_{2w}/FLAIR hyperintense peri-lesional compartment. A total of four experts were asked to perform the manual annotations. The senior-most expert (V.H expert 1, >10-years of experience in neuroradiology) independently annotated half of the studies, while expert 2 (V.S) with 7 years of experience in neuroradiology supervised expert 3 (K.B, with >3 years of radiology experience, and G. S. with >3 years of experience), to manually annotate the remaining cases individually. In rare cases with disagreement across the readers (expert 2, expert 3, and expert 4), the senior-most radiologist (V.H, expert 1) was consulted to reach consensus and obtain the final segmentations.

In order to map all scans to the same space for the purpose of spatial atlas construction, the Gd-T_{1w} MRI sequence of each patient was co-registered to a healthy 1.0-mm isotropic T1-weighted brain atlas (MNI152; Montreal Neurological Institute), using mutual-information-based similarity measure provided in ANTs (Advanced Normalization Tools) SyN (Symmetric Normalization) toolbox (Avants et al., 2008). This toolbox was employed due to its proved efficiency in mapping brain images containing lesions into healthy templates (Eloyan et al., 2014). In order to ensure exclusion of intensity differences within the tumor regions while only considering intensity differences from healthy tissue, the entire tumor mask was removed during registration. Skull stripping was then performed using a deformable surface classification algorithm (Tao and Chang, 2010), followed by bias field correction that was performed using the non-parametric non-uniform intensity normalization technique in Tustison et al. (2010).

Frequency Map Construction

From the available annotations for both enhancing lesion and T_{2w}/FLAIR hyperintense peri-lesional compartments, population atlases for each compartment were built for both pathologies (tumor recurrence and PsP). These atlases were constructed to quantify the frequency of occurrence of both enhancing lesion and peri-lesional hyperintensities across tumor recurrence and PsP, by averaging intensity values for all voxels across all the annotated binary images of all patients involved in the study. The frequency of lesion occurrence was visualized using a heat map superimposed on the reference MNI152 atlas.

Analysis of Differential Involvement (ADDIFI)

From the constructed tumor progression and PsP frequency atlases, analysis of differential involvement (ADIFFI) was performed as described in Ellingson et al. (2012), once for the enhancing lesion compartment and once for the peri-lesional hyperintensities. ADIFFI has been previously applied and shown success in the literature in the context of similar clinical problems (Ellingson et al., 2012; Kinoshita et al., 2014). ADIFFI employs Fisher's exact test on a voxel-wise basis, where the test yields exact p-values based on contingency tables (McDonald, 2009). Fisher's exact test is also recommended in the cases with two nominal variables, where there is a need to assess whether the proportions of one variable are different depending on the value of the other variable (McDonald, 2009).

First, a two-tailed Fisher's exact test was conducted, to evaluate a 2 x 2 contingency table that compares tumor recurrence/PsP along with tumor/non-tumor occurrence for each voxel across all patients. From this voxel-wise analysis, significance level was then measured, and the voxels that yielded p-value < 0.05 were stored. The voxel-wise probabilities according to Fisher's exact test are computed using the following formula:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!},$$

where a, b, c, d, and n are defined as follows:

- a: represents the number of tumor recurrence as well as the lesion-positive occurrences across all subjects at the current voxel.
- b: represents the number of tumor progression as well as the lesion-negative occurrences across all subjects at the current voxel.
- c: represents the number of PsP as well as the lesion-positive occurrences across all subjects at the current voxel.
- d: represents the number of PsP as well as the lesion-negative occurrences across all subjects at the current voxel.
- n: represents the total number of studies.

Next, connected component analysis was applied (Vincent, 1993), to cluster all statistically significant voxels found across the two compartments for both tumor recurrence and PsP that appeared on the ADIFFI maps, for enhancing lesion as well as for peri-lesional hyperintensities. The brain was finally partitioned using pre-labeled anatomical structures in MNI space (Mazziotta et al., 2001), for the purpose of identifying the anatomic areas of localization for tumor recurrence/PsP across all subjects.

Cluster-Size Correction Using Random Permutation Analysis

Due to the extensive number of voxel-wise calculations performed during ADIFFI, multiple comparison corrections were performed, which also aim at isolating the spatially distinct clusters associated with significant differences between the two groups. Random permutation (RP) analysis was conducted for

cluster size correction (Bullmore et al., 1999). Specifically, our task was to determine to what extent we can randomly obtain a cluster of statistical significance the same size or larger than the observed pattern in the original tumor recurrence versus PsP statistical maps. In order to achieve this, all T_{2w}/FLAIR hyperintense peri-lesional components, as well as the enhancing lesion ones, across the two categories (tumor recurrence/ PsP) were randomly reassigned to one of these pathologies, then ADIFFI was re-conducted, and voxels with p-values < 0.05 were stored. In addition, the sizes of statistically significant clusters were documented at each iteration. The whole process was reiterated across 500 iterations. RP analysis was employed in order to identify distinct clusters occurring <5% by chance, which would provide distinct spatial differences between tumor recurrence and PsP.

Finally, statistically significant clusters appearing on the cluster-size corrected ADIFFI maps were designated as either PsP or tumor recurrence by referring to the population atlases that were individually constructed for tumor recurrence and PsP. A specific anatomic localization was then obtained from these cluster-size corrected ADIFFI maps, by mapping them to a structural MNI atlas. The entire pipeline of this work is shown in **Figure 1**.

RESULTS

The resulting frequency maps that were constructed for both T2w/FLAIR hyperintense peri-lesional and lesion areas from pre-treatment scans are shown in **Figures 2**, **3**, respectively. These figures show that tumor recurrence in both compartments (enhancing lesion and T2w/FLAIR hyperintense peri-lesional areas) is more likely lateralized toward the parietal lobe, whereas PsP is more likely to be multi-focally distributed across different anatomical areas of the brain including frontal and temporal lobes, the insula, and the putamen.

Tumor Recurrence Is Lateralized Toward the Parietal Lobe

The frequency maps as well as ADIFFI maps for peri-lesional T2/FLAIR hyperintensities of the pre-treatment scans show that tumor recurrence is more likely to be present in the parietal lobe, with frequency of occurrence of 85% (59% of this distribution was found in the right hemisphere, whereas 41% was found in the left hemisphere), 13% in the occipital lobe (83% in the right hemisphere and 17% in the left hemisphere), and 2% in the right temporal lobe (**Figures 2A**, **4A**). Frequency maps as well as ADIFFI maps obtained for the enhancing lesion also reveal that tumor recurrence is more likely to be present in the parietal lobe of left and right hemispheres (70% and 30% chances of occurrence, respectively), **Figures 3A**, **4C**. These results suggest that tumor recurrence exhibits lobar prominence across the population atlases, but do not exhibit any hemisphere-specific preference. These lobar percentages were obtained by parcellating the brain with respect to an MNI structural atlas, shown in **Figure 4E**.

Pseudo-Progression Exhibits a Multi-Focal Distribution in the Enhancing Lesion as Well as the Perilesional Hyperintensities

PsP, unlike tumor recurrence, seems to more likely be multi-focally distributed across the brain in pre-treatment cases, for both the enhancing lesion and the peri-lesional hyperintensities. PsP exhibited a multi-focal distribution in the right hemisphere of the peri-lesional hyperintensities, with frequencies of occurrence of 55% in the frontal lobe, 11% in the temporal lobe, 10% in the insula, 10% in the putamen, and 9% in the parietal lobe (77% in the right hemisphere and 23% in the left hemisphere), and 5% in the right thalamus (**Figures 2B**, **4B**). In the analysis of the enhancing lesion regions, PsP appears to more likely be multi-focally distributed within both left and right hemispheres. The spatial distribution was 35% in the insula (with 63% of this distribution in the right hemisphere and 37% in the left hemisphere), 21% in the right frontal lobe, 13% in the right temporal lobe, 17% in the putamen (with 57% of this distribution in the right hemisphere and 43% in the left hemisphere), and 14% in the right parietal lobe (**Figures 3B**, **4D**).

Random Permutation Analysis for Cluster Size Correction

RP analysis conducted on the peri-lesional T2/FLAIR hyperintensities of the pre-treatment cases revealed that the average and standard deviation of maximum cluster size are 3,700 and 1726.8 voxels, respectively. Also, 95% of the cluster sizes were smaller than 6,192 voxels, meaning that clusters larger than this size threshold would occur in <5% of all random permutations. This resulted in one distinct T2w/FLAIR hyperintense peri-lesional cluster size of 6,502 voxels, localized at the right parietal lobe, and associated with tumor recurrence, and another one of size of 6,200 voxels localized at the left parietal lobe.

RP analysis conducted on the enhancing lesion revealed that average and standard deviation of maximum cluster size are 2,258 and 1774.1 voxels, respectively. Also, 95% of the cluster sizes were smaller than 5,164 voxels, meaning that clusters larger than this size threshold would occur in <5% of all random permutations. This resulted in one distinct enhancing lesion cluster size of 5,450 voxels, localized at the left parietal lobe, and associated with tumor recurrence.

The designation of PsP or true progression based on ADIFFI maps as for each significant voxel/cluster was accomplished by referring to the population atlases of both compartments (enhancing lesion, T2w/FLAIR hyperintense peri-lesion) that were individually constructed for tumor recurrence and PsP. The cluster-size corrected ADIFFI maps obtained for tumor recurrence are shown in **Figure 1D**.

Apart from the probabilistic approach conducted above, we conducted a statistical experiment, where we used a two-sample t-test, that was performed on the clinical parameters of tumor recurrence versus pseudo-progression, namely extent of resection, age, and gender, to obtain the 5% significance level.

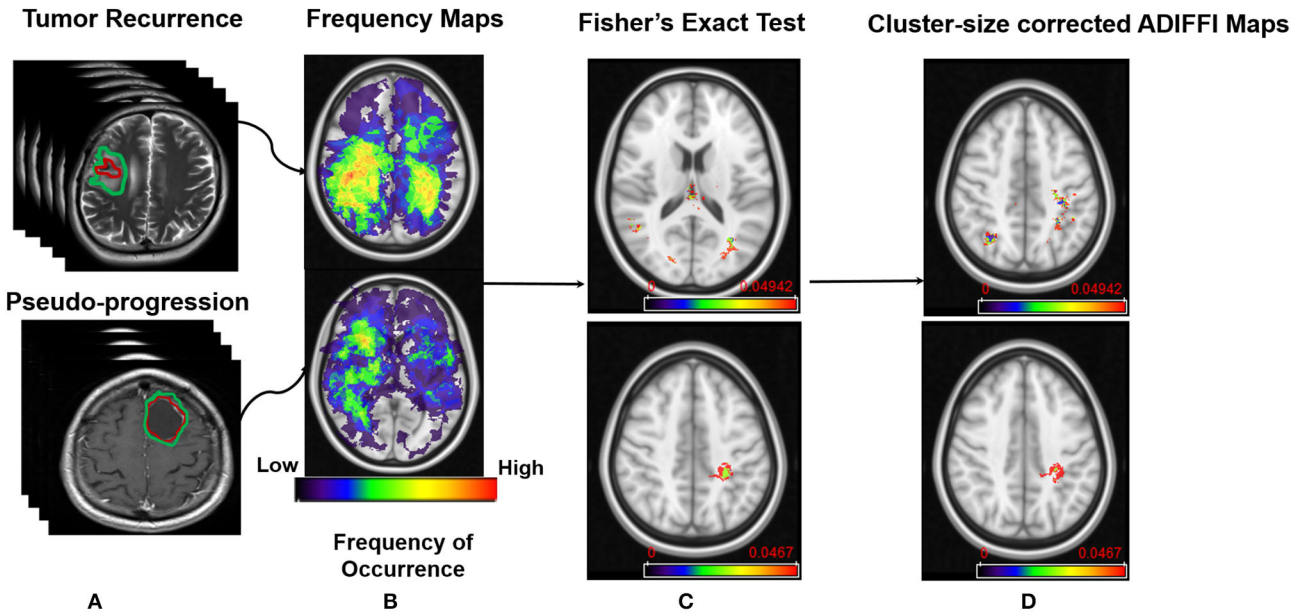


FIGURE 1 | Pipeline of the framework. **(A)** MR scans of tumor recurrence and pseudo-progression. **(B)** Frequency map atlases that were constructed from the two classes. **(C)** Results from Fisher's Exact test on peri-lesional T2/FLAIR hyperintensities in tumor recurrence (Top), and enhancing lesion in tumor recurrence (Bottom). **(D)** Results after applying RP analysis on ADIFFI maps shown in **(C)**.

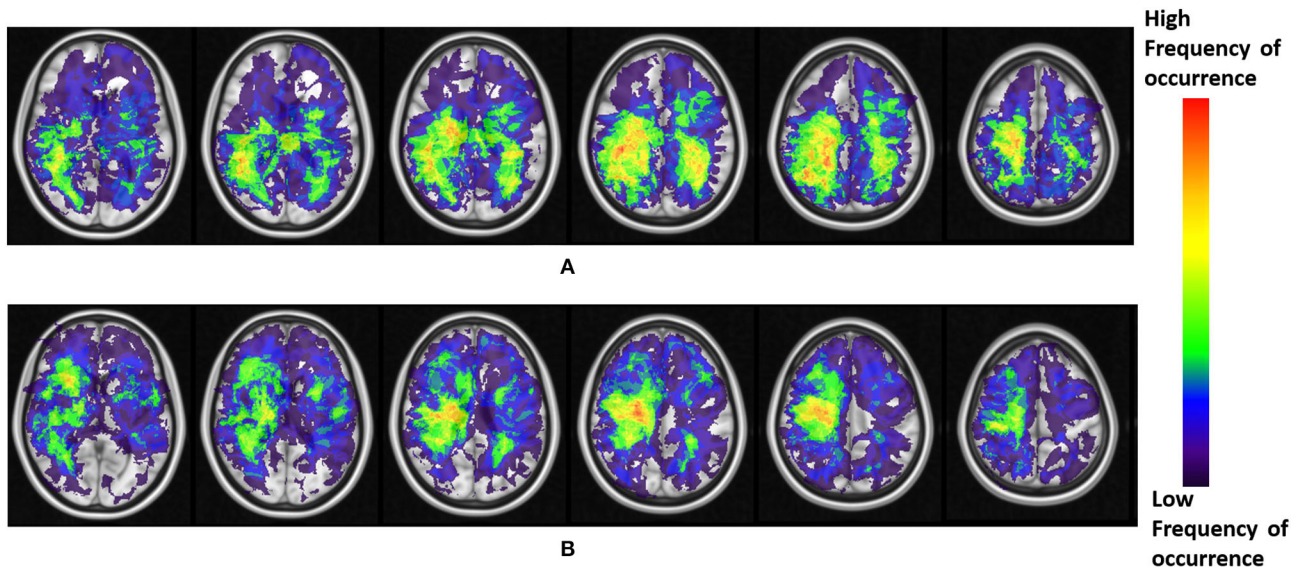


FIGURE 2 | (A) Frequency maps of tumor occurrence for peri-lesional T2/FLAIR hyperintensities in tumor recurrence of pre-treatment scans, where lobar prominence is present in the parietal lobe of both hemispheres. **(B)** Frequency maps of tumor occurrence for peri-lesional T2/FLAIR hyperintensities in pseudo-progression, where a multi-focal distribution is present in the frontal lobe, temporal lobe, insula, and putamen of the right hemisphere.

We found that the difference between the two pathologies was not statistically significant, based on each of the 3 parameters, p-values $= 0.52$, 0.25, and 0.82 for extent of resection, age, and gender, respectively. All of this information is available in **Supplementary Material**.

DISCUSSION

Distinguishing tumor recurrence from PsP is one of the biggest clinical challenges in GBM management. This feasibility study aimed at creating population atlases to study spatial proclivity

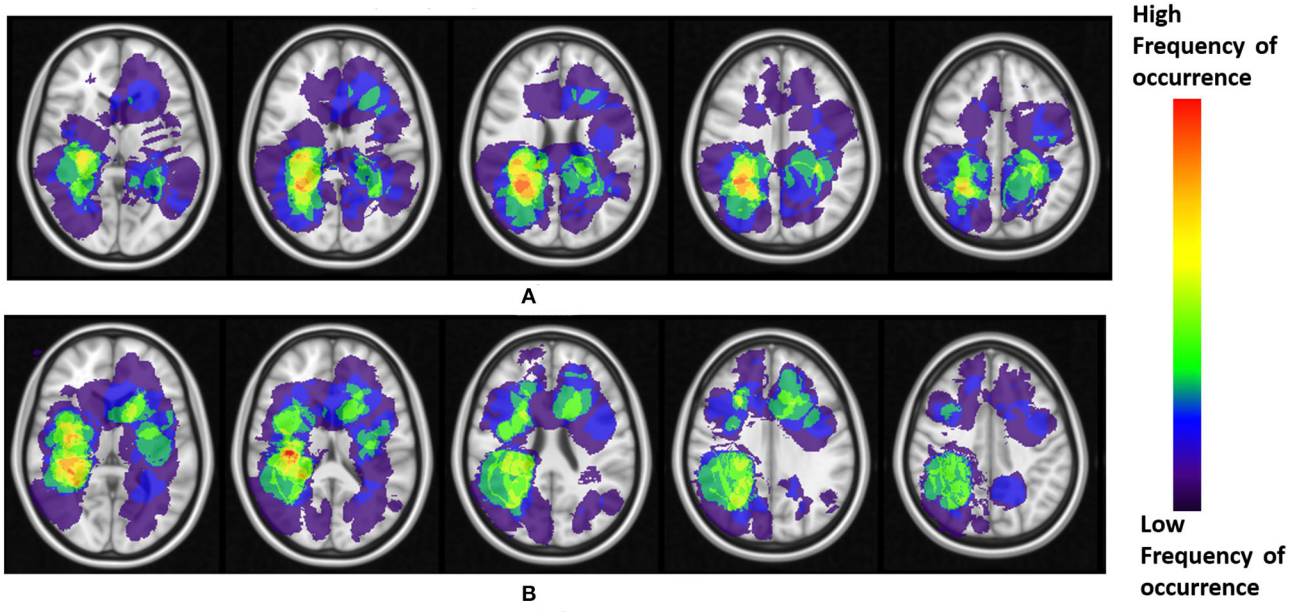


FIGURE 3 | (A) Frequency maps of tumor occurrence for enhancing lesion in tumor recurrence of pre-treatment scans, where lobar prominence is present in the parietal lobe of both hemispheres. **(B)** Frequency maps of tumor occurrence for enhancing lesion in pseudo-progression, where a multi-focal distribution is present in the insula, frontal lobe, putamen, and the temporal lobe, of both left and right hemispheres.

of brain tumor recurrence vs. PsP based on their occurrences on pre-treatment MR scans. The study assessed the voxel-wise tumor frequency across two lesion compartments using a statistical mapping technique named ADIFFI, in efforts to find significant spatial distribution differences between the two phenotypes.

Our preliminary findings suggest that likelihood of tumor recurrence is more consistent with lesions occurring in the parietal lobe of both left and right hemispheres, based on the analysis of both enhancing lesion and peri-lesional T2/FLAIR hyperintensities, on pre-treatment MRI scans. Parietal lobe is largely responsible for cognitive functions. Damage to parietal lobe may have direct implications in processing speech as well as sensory information. Hence, presence of tumor recurrence in parietal lobe may cause symptoms associated with numbness and tingling, hemi-neglect, and cognitive issues around right-left confusion and reading and math problems. PsP, on the other hand, did not exhibit lobar-specific distribution in pre-treatment scans, but showed a multi-focal distribution of the initial tumor in the frontal (associated with motor function, memory, problem solving) and temporal lobes (associated with primary auditory perception, such as hearing and visual recognition) as well as the insula and putamen. While the association of presence of tumor recurrence or PsP with specific lobes in the brain is not well-understood, their presence in specific lobes could ultimately contribute toward making more informed decisions regarding their diagnosis.

Previous studies have largely employed population atlases in brain tumors using pre-treatment MRI to obtain probabilistic maps of spatial predisposition in patients based on their disease aggressiveness (Duffau and Capelle, 2004) or molecular status (Drabycz et al., 2010; Ellingson et al., 2012; Kanas et al., 2017). For instance, a few studies have shown that tumor recurrence closer to the ventricular system was significantly associated with poor survival (Jafri et al., 2012; Adeberg et al., 2014). Interestingly, the study in Liu et al. (2016) showed that tumors in the right occipito-temporal periventricular white matter were significantly associated with poor survival in both training and test cohorts. Similarly, more aggressive GBMs were reported to be close to the ventricular system, and had a rapid progression (Li et al., 2018), suggesting that tumor location may play a significant role in disease etiology.

The closest studies to our work have attempted to identify associations of lesion location with likelihood of tumor recurrence and PsP, to investigate any spatial differences between the two phenotypes. For instance, the study by Tsien et al. (2010) incorporated location along with clinical and conventional MRI parameters to distinguish tumor progression from PsP in high-grade gliomas, yet no significant location differences could be found between the two groups, perhaps on account of the relatively small population size involved in this study (27 patients total). The study by Van West et al. (2017) reported the incidence of PsP in low grade gliomas, and found that 50% of their PsP enhancing lesions were located in the periventricular walls; attributing to the relatively poor blood supply in the periventricular areas that make it more vulnerable to radiation-induced processes. However, these studies did not report any findings regarding lobular preferences for either PsP or tumor recurrence in GBMs.

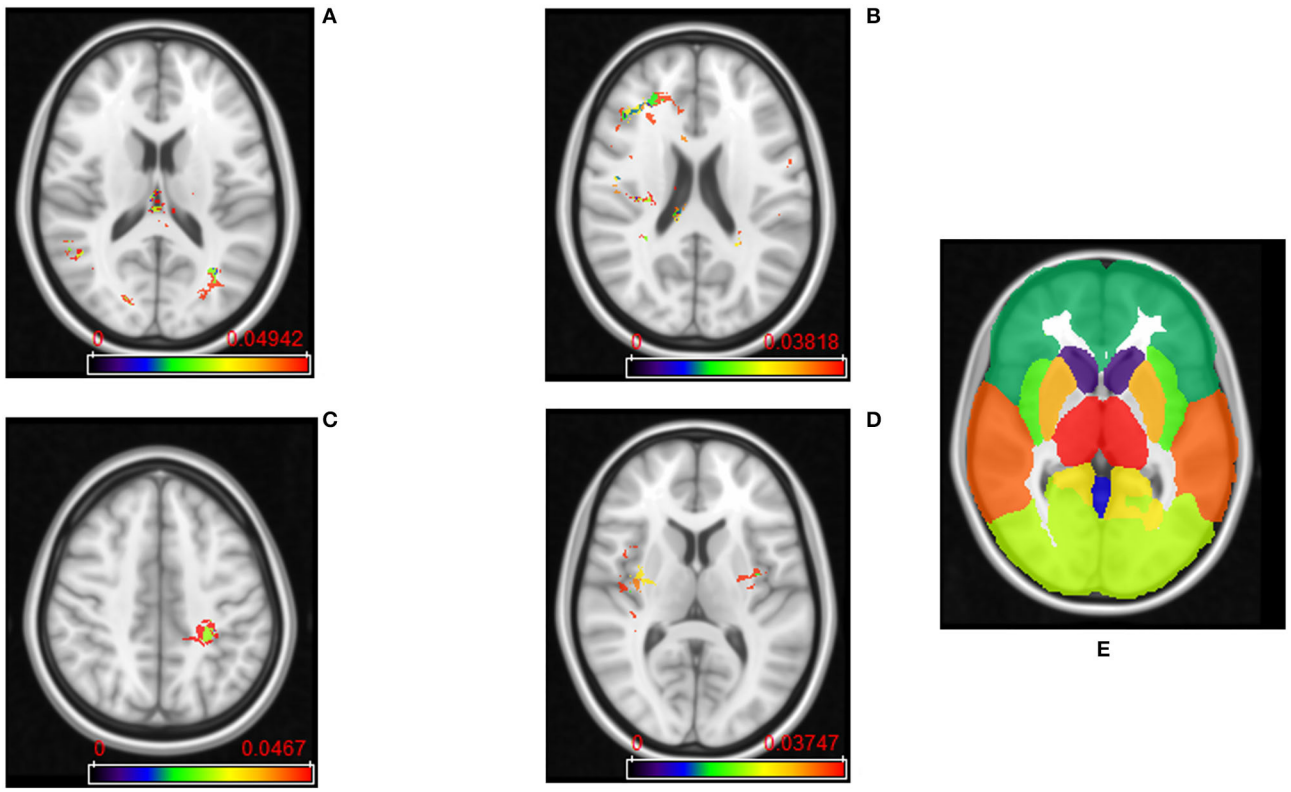


FIGURE 4 | (A) ADIFFI maps for peri-lesional T2/FLAIR hyperintensities in tumor recurrence, and **(B)** pseudo-progression. **(C)** ADIFFI maps for enhancing lesion in tumor recurrence, and **(D)** pseudo-progression. The level of significance was at a *p*-value of 0.05 for all of these maps. These were the maps prior to applying RP analysis. **(E)** The labeled anatomical MNI atlas that is used for parcellating ADIFFI maps and identifying significant areas.

Our study did have its limitations. First, our dataset is relatively small (74 studies). However, our sample size of $n = 74$ studies is comparable to existing studies in the literature on distinguishing PsP from tumor recurrence with sample sizes ranging from $n = 19$ to $n = 98$ (Cha et al., 2014; Wang et al., 2016; Boxerman et al., 2017; Elshafeey et al., 2019). Additionally, our work, similar to some of the published studies in distinguishing PsP vs. tumor recurrence (Cha et al., 2014; Elshafeey et al., 2019), did not include a separate hold-out validation cohort for analysis. Future work will focus on obtaining additional pre-treatment cases to further investigate our spatial predisposition findings, for tumor recurrence and pseudo-progression on large multi-institutional studies, as well as validate our findings on a separate independent patient cohort. In addition, while our results are promising as a feasibility study, our study did not account for molecular status (i.e., MGMT), or Karnofsky performance score as potential confounders during analysis. A potential limitation of this study is the lack of advanced imaging modalities such as dynamic susceptibility contrast (DSC), and Fluoro-O-(2) fluoroethyl-l-tyrosine (FET), which could have allowed for a joint multi-modal analysis combining these modalities with the probabilistic atlases. Additionally, one of our future directions includes extensively evaluating different automated segmentation approaches on the tumor compartments for the constructed probabilistic atlases, to extend our feasibility analysis. We also plan to obtain multiple segmentations from different readers for every study, to assess the impact of segmentation variability on our analysis. The prognostic implications (i.e., predicting patient overall survival), based on the location differences across PsP and tumor recurrence will also be investigated in the future.

To conclude, this study attempted to demonstrate the likelihood of occurrence of tumor recurrence and pseudo-progression, using the location of the lesion on pre-treatment MR scans. Our results revealed distinct localization between tumor recurrences and PsP that could aid in predicting these two similar appearing pathological conditions. Future work will focus on integrating the location biomarker with other biomarkers, such as shape and texture features, on a larger cohort of multi-institutional studies. We will also consider identifying location specific markers associated with radiation necrosis (delayed treatment effects) vs. tumor recurrence.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Cleveland Clinic Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MI and PT contributed to analysis and interpretation of data and contributed to designing the experiments, drafting, and revising the article. VH, VS, and EM provided clinical datasets and interpretation of radiographic images. MA helped define the clinical problem and provided clinical interpretation of findings. VH, RC, PP, GS, KB, and RT curated the studies and performed the annotations on radiological images. AM revised the manuscript critically for important intellectual content. All authors have reviewed the manuscript.

FUNDING

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, and R01 CA220581-01A1, National Center for Research Resources under award number 1 C06 RR12463-01, the DOD Prostate Cancer Idea Development Award; the DOD Lung Cancer Idea Development Award; Dana Foundation David Mahoney Neuroimaging Program, the Ohio Third Frontier Technology Validation Fund, the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering and the Clinical and Translational Science Award Program (CTSA) at Case Western Reserve University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Benjamin Ellingson for his guidance throughout the statistical experiment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.563439/full#supplementary-material

REFERENCES

Adeberg, S., König, L., Bostel, T., Harrabi, S., Welzel, T., Debus, J., et al. (2014). Glioblastoma recurrence patterns after radiation therapy with regard to the subventricular zone. *Int. J. Radiat. Oncol. Biol. Phys.* 90, 886–893. doi: 10.1016/j.ijrobp.2014.07.027

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004

Bilello, M., Akbari, H., Da, X., Pisapia, J. M., Mohan, S., Wolf, R. L., et al. (2016). Population-based MRI atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma. *NeuroImage* 12, 34–40. doi: 10.1016/j.nicl.2016.03.007

Boxerman, J. L., Ellingson, B. M., Jeyapalan, S., Elinzano, H., Harris, R. J., Rogg, J. M., et al. (2017). Longitudinal DSC-MRI for distinguishing tumor recurrence from pseudoprogression in patients with a high-grade glioma. *Am. J. Clin. Oncol.* 40, 228–234. doi: 10.1097/COC.0000000000000156

Brandsma, D., Stalpers, L., Taal, W., Sminia, P., and van den Bent, M. J. (2008). Clinical features, mechanisms, and management of pseudo-progression in malignant gliomas. *Lancet Oncol.* 9, 453–461. doi: 10.1016/S1470-2045(08)70125-6

Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., and Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imaging* 18, 32–42. doi: 10.1109/42.750253

Cha, J., Kim, S. T., Kim, H. J., Kim, B. J., Kim, Y. K., Lee, J. Y., et al. (2014). Differentiation of tumor progression from pseudoprogression in patients with posttreatment glioblastoma using multiparametric histogram analysis. *Am. J. Neuroradiol.* 35, 1309–1317. doi: 10.3174/ajnr.A3876

Chuang, M. T., Liu, Y. S., Tsai, Y. S., Chen, Y. C., and Wang, C. K. (2016). Differentiating radiation-induced necrosis from recurrent brain tumor using MR perfusion and spectroscopy: a meta-analysis. *PloS ONE* 11:e0141438. doi: 10.1371/journal.pone.0141438

Detsky, J. S., Keith, J., Conklin, J., Symons, S., Myrehaug, S., Sahgal, A., et al. (2017). Differentiating radiation necrosis from tumor progression in brain metastases treated with stereotactic radiotherapy: utility of intravoxel incoherent motion perfusion MRI and correlation with histopathology. *J. Neurooncol.* 134, 433–441. doi: 10.1007/s11060-017-2545-2

Drabycz, S., Roldán, G., De Robles, P., Adler, D., McIntyre, J. B., Magliocco, A. M., et al. (2010). An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging. *Neuroimage* 49, 1398–405. doi: 10.1016/j.neuroimage.2009.09.049

Duffau, H., and Capelle, L. (2004). Preferential brain locations of low-grade gliomas: comparison with glioblastomas and review of hypothesis. *Cancer* 100, 2622–2626. doi: 10.1002/cncr.20297

Ellingson, B. M., Cloughesy, T. F., Pope, W. B., Zaw, T. M., Phillips, H., Lalezari, S., et al. (2012). Anatomic localization of O6-methylguanine DNA methyltransferase (MGMT) promoter methylated and unmethylated tumors: a radiographic study in 358 *de novo* human glioblastomas. *Neuroimage* 59, 908–916. doi: 10.1016/j.neuroimage.2011.09.076

Eloyan, A., Shou, H., Shinohara, R. T., Sweeney, E. M., Nebel, M. B., Cuzzocreo, J. L., et al. (2014). Health effects of lesion localization in multiple sclerosis: spatial registration and confounding adjustment. *PloS ONE* 9:e107263. doi: 10.1371/journal.pone.0107263

Elshafeey, N., Kotrotsou, A., Hassan, A., Elshafei, N., Hassan, I., Ahmed, S., et al. (2019). Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudo-progression in glioblastoma. *Nat. Commun.* 10, 1–9. doi: 10.1038/s41467-019-11007-0

Jafri, N. F., Clarke, J. L., Weinberg, V., Barani, I. J., and Cha, S. (2012). Relationship of glioblastoma multiforme to the subventricular zone is associated with survival. *Neuro-oncology* 15, 91–96. doi: 10.1093/neuonc/nos268

Kanas, V. G., Zacharaki, E. I., Thomas, G. A., Zinn, P. O., Megalooikonomou, V., and Colen, R. R. (2017). Learning MRI-based classification models for MGMT methylation status prediction in glioblastoma. *Comput. Methods Programs Biomed.* 140, 249–257. doi: 10.1016/j.cmpb.2016.12.018

Kikinis, R., Pieper, S. D., and Vosburgh, K. G. (2014). "3D slicer: a platform for subject-specific image analysis, visualization, and clinical support," in *Intraoperative Imaging and Image-Guided Therapy,* eds F. Jolesz (New York, NY: Springer), 277–289. doi: 10.1007/978-1-4614-7657-3_19

Kinoshita, M., Sasayama, T., Narita, Y., Yamashita, F., Kawaguchi, A., Chiba, Y., et al. (2014). Different spatial distribution between germinal center

B and non-germinal center B primary central nervous system lymphoma revealed by magnetic resonance group analysis. *Neuro-oncology* 16, 728–734. doi: 10.1093/neuonc/not319

Laigle-Donadey, F., Martin-Duverneuil, N., Lejeune, J., Criniere, E., Capelle, L., Duffau, H., et al. (2004). Correlations between molecular profile and radiologic pattern in oligodendroglial tumors. *Neurology* 63, 2360–2362. doi: 10.1212/01.WNL.0000148642.26985.68

Larjavaara, S., Mäntylä, R., Salminen, T., Haapasalo, H., Raitanen, J., Jääskeläinen, J., et al. (2007). Incidence of gliomas by anatomic location. *Neuro-oncology* 9, 319–325. doi: 10.1215/15228517-2007-016

Li, H. Y., Sun, C. R., He, M., Yin, L. C., Du, H. G., and Zhang, J. M. (2018). Correlation between tumor location and clinical properties of glioblastomas in frontal and temporal lobes. *World Neurosurg.* 112, 407–414. doi: 10.1016/j.wneu.2018.01.055

Liu, T. T., Achrol, A. S., Mitchell, L. A., Du, W. A., Loya, J. J., Rodriguez, S. A., et al. (2016). Computational identification of tumor anatomic location associated with survival in 2 large cohorts of human primary glioblastomas. *Am. J. Neuroradiol.* 37, 621–28. doi: 10.3174/ajnr.A4631

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001). A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 356, 1293–1322. doi: 10.1098/rstb.2001.0915

McDonald, J. H. (2009). *Handbook of Biological Statistics*. Baltimore, MD: Sparky House Publishing

Metellus, P., Coulibaly, B., Colin, C., de Paula, A. M., Vasiljevic, A., Taieb, D., et al. (2010). Absence of IDH mutation identifies a novel radiologic and molecular subtype of WHO grade II gliomas with dismal prognosis. *Acta Neuropathol.* 120, 719–729. doi: 10.1007/s00401-010-0777-8

Parvez, K., Parvez, A., and Zadeh, G. (2014). The diagnosis and treatment of pseudo-progression, radiation necrosis and brain tumor recurrence. *Int. J. Mol. Sci.* 15, 11832–11846. doi: 10.3390/ijms150711832

Patel, U., Patel, A., Cobb, C., Benkers, T., and Vermeulen, S. (2014). The management of brain necrosis as a result of SRS treatment for intra-cranial tumors. *Transl. Cancer Res.* 3, 373–382. doi: 10.3978/j.issn.2218-676X.2014.07.05

Prager, A. J., Martinez, N., Beal, K., Omuro, A., Zhang, Z., and Young, R. J. (2015). Diffusion and perfusion MRI to differentiate treatment-related changes including pseudo-progression from recurrent tumors in high-grade gliomas with histopathologic evidence. *Am. J. Neuroradiol.* 36, 877–885. doi: 10.3174/ajnr.A4218

Stockhammer, F., Misch, M., Helms, H. J., Lengler, U., Prall, F., Von Deimling, A., et al. (2012). IDH1/2 mutations in WHO grade II astrocytomas associated with localization and seizure as the initial symptom. *Seizure* 21, 194–197. doi: 10.1016/j.seizure.2011.12.007

Tao, X., and Chang, M. C. (2010). A skull stripping method using deformable surface and tissue classification. *Proc. SPIE* 7623:76233L. doi: 10.1117/12.844061

Tsien, C., Galbán, C. J., Chenevert, T. L., Johnson, T. D., Hamstra, D. A., Sundgren, P. C., et al. (2010). Parametric response map as an imaging biomarker to distinguish progression from pseudo-progression in high-grade glioma. *J. Clin. Oncol.* 28:2293. doi: 10.1200/JCO.2009.25.3971

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Van West, S. E., de Bruin, H. G., van de Langerijt, B., Swaak-Kragten, A. T., Van Den Bent, M. J., and Taal, W. (2017). Incidence of pseudo-progression in low-grade gliomas treated with radiotherapy. *Neuro-oncology* 19, 719–725. doi: 10.1093/neuonc/now194

Vincent, L. (1993). Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *IEEE Trans. Image Process.* 2, 176–201. doi: 10.1109/83.217222

Wang, S., Martinez-Lage, M., Sakai, Y., Chawla, S., Kim, S. G., Alonso-Basanta, M., et al. (2016). Differentiating tumor progression from pseudoprogression in patients with glioblastomas using diffusion tensor imaging and dynamic susceptibility contrast MRI. *Am. J. Neuroradiol.* 37, 28–36. doi: 10.3174/ajnr.A4474

Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Gregory Sorensen, A., DeGroot, E. G., et al. (2010) Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J. Clin. Oncol.* 28, 1963–1972. doi: 10.1200/JCO.2009.26.3541

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 27 January 2021
doi: 10.3389/fncom.2020.495075

Improvement of Multiparametric MR Image Segmentation by Augmenting the Data With Generative Adversarial Networks for Glioma Patients

Eric Nathan Carver[1,2], *Zhenzhen Dai*[1], *Evan Liang*[1], *James Snyder*[1] *and Ning Wen*[1]*

[1] Henry Ford Health System, Detroit, MI, United States, [2] Wayne State University, Detroit, MI, United States

Every year thousands of patients are diagnosed with a glioma, a type of malignant brain tumor. MRI plays an essential role in the diagnosis and treatment assessment of these patients. Neural networks show great potential to aid physicians in the medical image analysis. This study investigated the creation of synthetic brain T1-weighted (T1), post-contrast T1-weighted (T1CE), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (Flair) MR images. These synthetic MR (synMR) images were assessed quantitatively with four metrics. The synMR images were also assessed qualitatively by an authoring physician with notions that synMR possessed realism in its portrayal of structural boundaries but struggled to accurately depict tumor heterogeneity. Additionally, this study investigated the synMR images created by generative adversarial network (GAN) to overcome the lack of annotated medical image data in training U-Nets to segment enhancing tumor, whole tumor, and tumor core regions on gliomas. Multiple two-dimensional (2D) U-Nets were trained with original BraTS data and differing subsets of the synMR images. Dice similarity coefficient (DSC) was used as the loss function during training as well a quantitative metric. Additionally, Hausdorff Distance 95% CI (HD) was used to judge the quality of the contours created by these U-Nets. The model performance was improved in both DSC and HD when incorporating synMR in the training set. In summary, this study showed the ability to generate high quality Flair, T2, T1, and T1CE synMR images using GAN. Using synMR images showed encouraging results to improve the U-Net segmentation performance and shows potential to address the scarcity of annotated medical images.

Keywords: GaN, U-net, glioma, GBM, segmentation

OPEN ACCESS

Edited by:
Spyridon Bakas, University of Pennsylvania, United States

Reviewed by:
Mohammad Hamghalam, Shenzhen University, China
Ujjwal Raghunandan Baid, Shri Guru Gobind Singhji Institute of Engineering and Technology, India

***Correspondence:**
Ning Wen
nwen1@hfhs.org

Received: *31 August 2019*
Accepted: *03 November 2020*
Published: *27 January 2021*

Citation:
Carver EN, Dai Z, Liang E, Snyder J and Wen N (2021) Improvement of Multiparametric MR Image Segmentation by Augmenting the Data With Generative Adversarial Networks for Glioma Patients. Front. Comput. Neurosci. 14:495075. doi: 10.3389/fncom.2020.495075

INTRODUCTION

Approximately 121,000 (Ostrom et al., 2018) people in the US are diagnosed with a malignant brain tumor annually, with over 13,000 of those being Glioblastoma (GBM), defined by the World Health Organization (WHO) as grade IV tumors with an unacceptable median overall survival despite best available treatment of less than to 2 years. For primary brain tumors WHO grade II-IV, there are no curative treatments and limited approved therapies. Current management of primary brain tumors has two standard benchmarks, tissue analysis for diagnosis, and the longitudinal analysis of treatment response/ tumor stability through serial brain tumor imaging.

In fact, the brain MRI in patients with GBM is used to stratify clinical trial options prior to initial surgery and to offer patients definitive cytoreduction surgery for malignant glioma or GBM when radiographic features are highly suggestive of a malignant tumor. Therefore, advanced imaging methods to stratify patients into phenotypic, functional, molecular, and prognostic groups is highly sought after.

Amongst GBM researchers, clinicians, patients, and patient advocates there is hope that new advances as promised by molecular targeted therapies, advanced radiation techniques, evolving surgical technologies, and unforeseen innovation will result in improved patient outcomes. One key element to all of these are MR images; both for diagnosis and longitudinal patient monitoring. Applications of deep machine learning in brain tumor imaging has the potential to transition from a subjective analysis to objective analysis and create a new set of tools to refine treatment options, improve care quality, and ultimately impact patient care. One critical limitation to achieving the success seen in non-medical imaging is the volume of data needed to power deep machine learning. It is common knowledge that deep learning techniques are highly powerful when there are numerous training samples. However, in the medical field, especially in clinical trials, where limited numbers of training samples are accessible, deep learning models are easily overfitting during the training stage and perform poorly in prediction (Shen et al., 2017). Besides, annotation of medical images is generally expensive, time-consuming, and requires highly trained clinicians. Therefore, data argument has been widely used to increase the original dataset to improve the performance of supervised learning. One possible solution to overcoming the limited brain tumor imaging data available for analysis is to create synthetic brain tumor MR images. Synthetic MR (synMR) images of sufficient quality may be created using a generative adversarial network (GAN). Herein, we quantitively and qualitatively evaluated the quality of these created synMR and established the capability of using synMR images for the practical application of increasing the volume of data required by deep learning. Specifically, we evaluated the performance of image segmentation using a widely implemented two-dimensional (2D) U-Net model (Ronneberger et al., 2015) by augmenting real patients' T1-weighted (T1), post-contrast T1-weighted (T1CE), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (Flair) MR images data with varying amount of synMR images. In fact, the investigation of the changes in accuracy of enhancing tumor (ET), whole tumor (WT), and tumor core (TC), also known as the non-enhancing necrotic region, for glioma patients when incorporating varying amounts of synMR images may be the most practically useful metric in judging both quality and real-world usability of T1, T1CE, T2, and Flair synMR.

METHOD

Patient Population

Data was obtained from the BraTS multimodal Brain Tumor Segmentation Challenge 2018 (Menze et al., 2015; Bakas et al., 2017a,b, 2018). Nineteen different institutions provided a total of 210 patients for training and 66 patients for validation. T1, T1CE, T2, and Flair MR images were provided for each patient. Provided ET/WT/TC contouring was performed by one to four clinicians and approved by neuro-oncologists.

Image Pre-processing

BraTS provided T1, T2, and Flair MRI that were rigidly registered with T1CE, resampled ($1 \times 1 \times 1$ mm^3) and skull stripped. In addition to the pre-processing performed by BraTS, this study performed normalization and padding of each 2D MRI slice from 240×240 to 256×256. To aid in data balance between tumor and unlabeled areas, the z dimension in the training dataset was cropped to 64 slices from original 155 slices. This served to decrease amount of unlabeled data present during training and increase focus on the tumor regions for data augmentation and segmentation. Data augmentation was done by flipping each slice left/right to decrease dependence on location as the brain exhibits marked symmetry across the sagittal plane. No cropping was performed on the validation dataset as all 155 slices were segmented during validation.

Generative Adversarial Neural Network

We developed an augmentation network to create synMR images as a new augmentation approach to aid in overcoming the well-known limitation of available annotated medical image data. We manipulated semantic label maps of lesions in real MR (rMR) images, e.g., changing lesion locations or types, and then transferred the new label to synMR image using the augmentation network. Compared to traditional augmentation methods such as affine transformation or cropping, which could not guarantee standard anatomical structures, this approach introduced new data augmentations by varying tumor sizes, shapes and locations while maintained the authentic morphologic structures of brain.

Architecture

Our augmentation network consisted of a generator (blue box in **Figure 1**) and two discriminators (red and yellow box in **Figure 1**). The generator was used to generate synMR images from sematic label maps which in turn were derived from rMR images. The semantic label map was composed of normal brain tissue and GBM tumor segments. GBM segments were further classified into the ET/WT/TC regions derived from T2 rMR. Three quarters, one half and a quarter of maximum pixel values of T2 were the three categorizing thresholds that were used to segment normal brain tissues. Five categories of segments were generated in total. The discriminators were used to distinguish between synMR images and rMR images.

Generator

The generator consists of several components *Ci-*„ with each operating at a different resolution. The semantic label (256×256) is down sampled to provide segmentation layout at the different resolutions ($wi \times hi$, $wi = hi$). The first component, *C0,* gets down-sampled semantic labels at the resolution of $w0 = h0 = 4$ as input and then it generates feature maps as an output for the next component. For components *C1* to *Cn*, feature maps

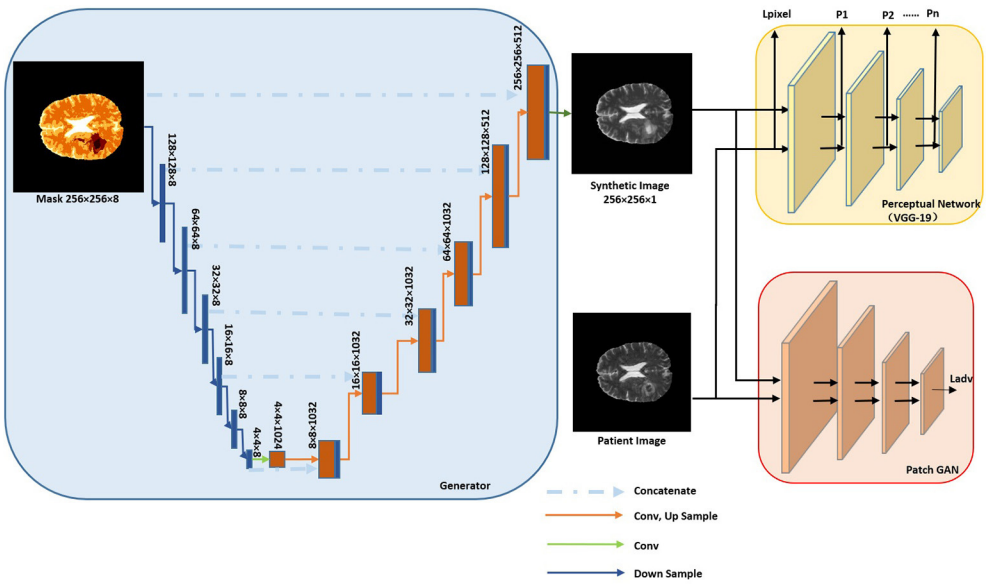


FIGURE 1 | Architecture of Augmentation Net. The Generator is shown in blue box and discriminators in yellow and red boxes.

from previous component are up sampled at a scale of 2 and are concatenated with the semantic labels of the same resolution as input. A residual block is applied to generate feature maps as output. The convolution kernel size is 3×3, layer normalization (Gomez-Iturriaga et al., 2016) is applied, and ReLU (Maas et al., 2013) is used as the activation function.

Discriminator

Two discriminators are used. The first one is a pre-trained VGG-19 convolutional neural network (Simonyan and Zisserman, 2014), which won the first and second place in localization and classification in the Image Net Large Scale Visual Recognition Challenge (ILSVRC) 2014. It is used to calculate the perceptual loss ($\sum_i p_i$) and the image per-pixel loss $\mathcal{L}_{im}$.

$$\mathcal{L}_{im} = \sum_m \sum_n |I_{real} - I_{synthetic}| \tag{1}$$

I_{real} represents the real patient rMR image, $I_{synthetic}$ represents the synMR image created by the generator, with $\sum_m \sum_n$ indicating the summation over all pixels

$$p_i = \sum_m \sum_n |\theta_{ireal} - \theta_{isynthetic}| \tag{2}$$

$p_i$is the perceptual loss from layer i of VGG-19 Net. Perceptual loss was firstly proposed by Johnson et al. (2016) and was claimed to be more robust than image per-pixel loss to measure image similarities. θ_{ireal} and $\theta_{isynthetic}$ are feature maps of rMR image and synMR images generated at layer i, respectively. The second one is a patch GAN, which penalizes on image patches, the loss is given as $\mathcal{L}_{adv} = \mathbb{E}\left[D\left(I_{real}, I_{synthetic}\right)\right] + \mathbb{E}\left[1 - D\left(I_{synthetic}\right)\right]$. $D(.)$ is the discriminator net. The total loss is computed as the weighted summation of each loss.

The synMR image is generated by solving the following objective:

$$S^* = argmin(\mathbb{E}\left[\sum_{i=0}^{n} \lambda_i p_i, + \lambda_{im}\mathcal{L}_{im}\right] + \lambda\mathcal{L}_{adv}) \tag{3}$$

Training and Generation of New Training Samples

One hundred sixty-four patients were randomly selected from the BraTS18 dataset for training. For each MRI modality, an independent model was trained. λ_i, λ_{im} and λ were adapted every 10 epochs to maintain the balance among each loss. The total training epoch was 100 for each modality. New semantic labels were created from real labels to augment synMR images. The lesion contours were rotated with a random angle (0°-90°), translated with a random number (0–40) of pixels and randomly flipped left/right/up/down. The lesion contours that were outside the brain contour were changed to zero (background). Then the augmented semantic labels were used as the new training dataset and transferred into image domain using the augmentation network of each imaging modality.

GAN Evaluation Metrics

Mean Square Error (MSE), Mean Absolute Error (MAE), Peak Signal to Noise Ratio (PSNR), and Structural Similarity Index (SSIM) were used to quantitatively compare between the synMR and rMR images.

In MSE (Equation 4) variable "n" represents number of images being compared. Since MSE depends on intensity scaling, it is necessary to report these details. In this study, 16-bit images were

used with the pixel range 0–255.

$$MSE = 1/\mathrm{n}\sum\{(original\ image - generated\ image)^2\} \quad (4)$$

MAE (Equation 5) determines the prediction error between the rMR and synMR.

$$MAE = 1/\mathrm{n}\sum\{(original\ image - generated\ image)\} \quad (5)$$

PSNR (Equation 6) overcomes the limitation of MSE by scaling the MSE value according to image range, which is done by the S^2 value in Equation 4. Generally, the higher the PSNR, the better the synthetic image; however, this metric has a limitation.

$$PSNR\ (dB) = -10{*}log_{10}\left(\frac{MSE}{S^2}\right) \quad (6)$$

SSIM shows the perceived change in structural information as opposed to MSE, MAE, and PSNR that show absolute error differences. SSIM assumes pixels close to each other possess strong inter-dependency. It is based on luminance, contrast, and structure differences between the images and is among the most commonly used metrics to compare the synthetic images to the original images.

The benchmark of quantitative metrics of our study deviated from other works on synthetic images as one of the key characteristics of our methodology was to produce variations of tumor size, shape and location in synMR. Therefore, there were inherent differences between the synMR and rMR which made direct quantitative comparisons difficult. To overcome this limitation, qualitative analysis of synMR images was performed in the form of the Turing test, physician individual synMR review, and investigation of changes in deep learning performance. The Turing test requires a physician to correctly classify a dataset consisting of both rMR and synMR. A misclassification percentage of fifty percent implies the rMR and synMR are indistinguishable. In addition to this test, an in-depth analysis of a randomly selected synMR was performed by an authoring physician. SynMR images were also assessed for the practical application of increasing the volume of data required by deep learning. Specifically, synMR was incorporated both in subgroups and as a whole during training of the outlined U-Net segmentation model. Investigation of impact on performance of segmentation could provide feedback on the quality of synMR images.

U-NET Segmentation Model

The segmentation model is comprised of three individual 2D U-Nets designed by Ronneberger et al. (2015), one for each of the three tumor regions: ET/WT/TC. Each U-Nets was trained with rMR and synMR images of modalities T1, T1CE, T2, and Flair. This model combines the ET/WT/ET contours generated by the three separate U-Nets during post-processing. Two processing techniques were used to improve the segmentation model's ability to accurately contour ET/TC. The first one served to aid the segmentation model during training by mathematically manipulating input T1 MRI to improve delineation of ET/TC boundaries. Specifically, each input T1 MRI was used in conjunction with its corresponding T1CE MRI and the pixel-wise intensity difference between these MRI was calculated. This calculated array replaced the T1 MRI during training. The second technique was to use the WT contour as a boundary for ET/TC delineations. Therefore, any ET/TC contour predicted outside of the WT contour would be erased. The best model for each type of contour was chosen according to the validation loss within 100 epochs run on GPU (Titan XP, nVidia, Santa Clara, CA).

U-NET Architecture

Each U-Net followed Pelt and Sethian (2018) recommendation of four downscaling and upscaling layers. Each downscaling layer is followed by a batch normalization layer (Pelt and Sethian, 2018) and the architecture uses this grouping to downsize the image while increasing the number of features. Each upscaling layer is merged with its corresponding downscale layer and used to return the downsized image to the size of the original. These layers combined to form a merged layer and soft dice (Equation 7) was employed as the loss function.

$$Dice\ Loss = \frac{2{*} < y_{true}, y_{pred} > + c}{< y_{true}, y_{true} > + < y_{pred}, y_{pred} > + c} \quad (7)$$

$y_{\text{-true}}$ is the clinician's contour, y_{pred} is the model's output, and c (0.01) is a constant to avoid division-by-zero singularities.

Creation of Training Datasets

As outlined previously, synMR images were generated from 210 GBM patients' rMR. To further investigate how synMR could affect segmentation performance during training of the U-Net, the synMR images were randomly partitioned into four unique subsets. Multiple U-Nets were trained using the total rMR in combination with each of these synMR subsets. One U-Net was trained using only rMR to serve as a baseline with which to compare performance. Four other U-Nets were trained on datasets that contained either a quarter, half, three-quarters or total generated synMR to investigate how the amount of synMR incorporated in the training dataset influences model performance. In order to solely evaluate the impact of amount of synMR images on the model performance, extra care was taken to decrease variance of the quality of synMR used in each training datasets. This was accomplished by dividing all synMR into four subsets equally, with each subset containing an exclusive quarter of all available synMR. These subsets were then numerically labeled one through four and used in the following manner to create the training datasets. Subset one was used to form the training dataset containing one quarter of synMR. To form the training dataset that employed half of the generated synMR, subset one was combined with subset two. Similarly, subsets one, two, and three were used to form training dataset representing three-quarters of available synMR, while all four subsets were used for the total synMR dataset. By staggering synMR subsets in each training model, we could evaluate the model performance differences with regards to change in the amount of synMR incorporated.

U-NET Evaluation Metrics

Dice similarity coefficient (DSC), Hausdorff distance with 95% confidence interval (HD), sensitivity, and specificity are used to evaluate U-Net segmentation as these metrics quantitatively show the agreement between the created U-Net model and the "gold-standard" physician created contours. Specifically, DSC indicates volumetric agreement of the physician created contour and the contours generated in this study. Reported DSC values fall into the range zero to one with zero indicating no volumetric overlap and one indicating complete volumetric agreement. HD indicates point-based agreement between the compared contours. This quantitative metric shows largest relevant Euclidean offset between every pixel in the ground truth contour and its corresponding pixel in the generated contour.

$$h(A,B) = max_{a\in A}\left\{min_{b\in B}\left\{d(a,b)\right\}\right\} \tag{8}$$

with *a* and *b* being points of sets *A* and *B*, respectively, and *d(a,b)* is the Euclidean metric between these points (Menze et al., 2015).

Sensitivity (true positive) and specificity (true negative) indicate level of border agreement between generated and physician contours. While DSC shows volumetric overlap of contours, these metrics report relative size differences. Essentially, they report if the generated contour is smaller or larger than the physician's contour.

$$Sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ +\ number\ of\ false\ negatives} \tag{9}$$

$$Specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives\ +\ number\ of\ false\ postives} \tag{10}$$

RESULTS

SynMR Quantitative Analysis

MSE, MAE, PSNR, and SSIM were performed to provide quantitative analysis of synMR. **Table 1** shows quantitative metric for each modality. Consistent inter-modality results demonstrate that high similarity is achieved between the rMR and synMR for all modalities (T1,T1CE,T2, Flair).

SynMR Qualitative Analysis

Qualitative analysis was performed to further investigate both overall and inter-modality synMR quality. Specifically, qualitative analysis of synMR was assessed in two ways by an authoring physician. The first assessment was the performance of the Turing test (**Table 2**). The second one was an in-depth visual comparison of generated synMR images with their corresponding rMR images on each of the following modalities: T2, Flair, T1, and T1CE.

SynMR Qualitative Analysis: Turing Test

A subset of 9 rMR images and 10 synMR images flair, T1, T1CE, and T2 MR images were randomly selected for evaluation. The physician was presented with each of these 19 MR images blindly and judged if the MR image was rMR or synMR and provided feedback. Ideally the rMR and synMR images would be completely indistinguishable from each other, and this would be reflected by a 50 percent misclassification rate of the images. As shown in **Table 2**, Flair, T1CE, and T2 MR images were misclassified 26.3 percent of the time, while T1 was incorrectly identified 10.5 percent of the time. This lower score was due to the visible streaking artifacts on coronal and sagittal views for some of the synMR images.

TABLE 1 | Average reported Mean Square Error (MSE), Mean Absolute Error (MAE), Peak Signal to Noise Ratio (PSNR), and Structural Similarity Index (SSIM) for synMR images generated by GAN.

	MSE	MAE	PSNR	SSIM
T1	19.3 ± 0.3	23.4 ± 0.6	43.1 ± 0.4	0.788 ± 0.002
T1CE	19.2 ± 0.3	22.8 ± 0.6	43.1 ± 0.4	0.789 ± 0.004
T2	19.2 ± 0.3	23.4 ± 0.4	43.1 ± 0.5	0.784 ± 0.003
Flair	18.9 ± 0.4	24.1 ± 1.5	43.1 ± 0.5	0.794 ± 0.005

TABLE 2 | The classification accuracy of a subset of synMR and original images reviewed by the physician blindly.

Modality	% Misclassified
Flair	26.3
T1	10.5
T1CE	26.3
T2	26.3

The amount of synMR and rMR improperly categorized by physician is represented by percent misclassified.

SynMR Qualitative Analysis: In-Depth Physician Analysis

Figure 2 shows that the T2 MR image's main difference between the synMR and rMR lay in the tumor at the right frontal lobe (lower left on the images A3 vs. A7). It was noted that the tumor geometry was preserved, but the relative signal intensities in the region were distorted. Specifically, synMR differed in appearance in the core of the tumor, as it displayed a hyperintense T2 signal compared with the surrounding edema. In addition, the signal from edema was also slightly different in images A4 and A8. Image A8 had a broader range of contrasts within the edema, whereas A4 delineated the extent of the edema with a sharper drop-off at the edges than the rMR. Comparing the edema between Flair rMR images and synMR images (B3 vs. B7), it showed differences in the extent of the edema. Also, there were noted circumferential artifacts in the rMR flair images (B1–B4). T1 MR images showed that quality of synMR (C1–C4) was very good. For T1CE MR images, the boundary of enhanced rim and necrotic regions of the T1CE synMR (D1–D4) were clearly defined, although the area surrounding the tumor had slight decrease in intensity.

In summary, synMR had high image quality with clearly defined structural boundaries. However, synMR suffered in showing details inside lesions and areas of high gradient (e.g., edema signal in T2 modalities). It was possible that this detailed

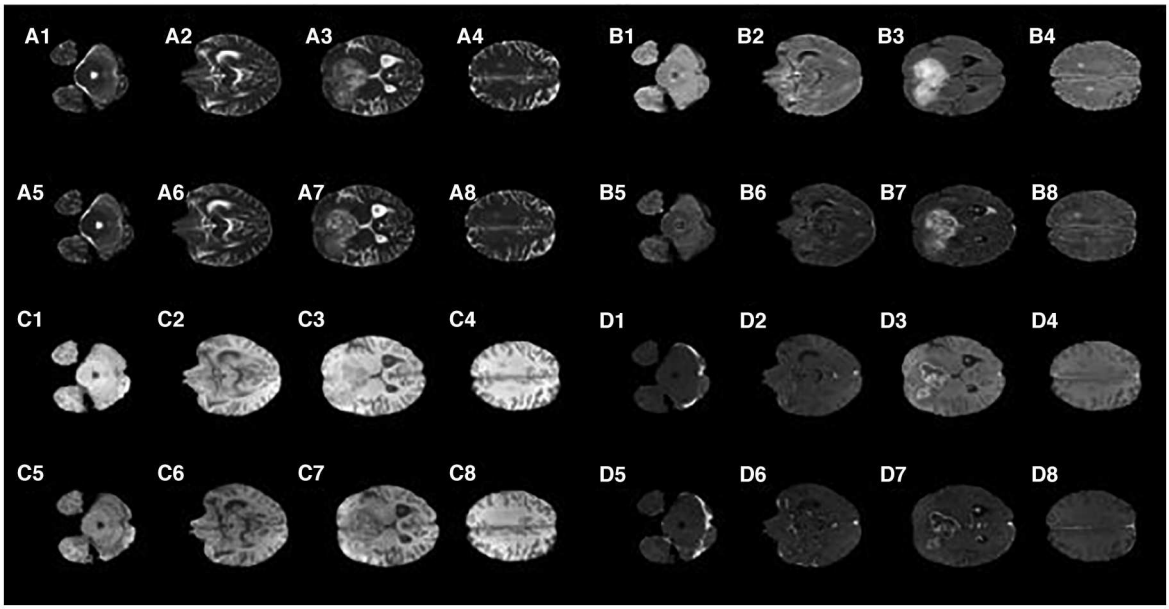


FIGURE 2 | Synthetic MRI compared with patient MR images. **(A1–A4)**, synthetic T2; **(A5–A8)**, patient T2; **(B1–B4)**, synthetic Flair; **(B5–B8)**, patient Flair; **(C1–C4)**, synthetic T1; **(C5–C8)**, patient T1; **(D1–D4)**, synthetic T1CE; **(D5–D8)**, patient T1CE.

information was lost when lesion pixels were classified into the same semantic label. This is notable as re-gaining lost information is a well-known GAN limitation.

U-Net: Utilization of SynMR

In addition to quantitative and qualitative investigation of synMR image quality, incorporation of generated synMR in training datasets for the U-Net segmentation model was done to assess the ability of synMR to enhance segmentation performance. The impact on the U-Net's segmentation performance by incorporating synMR during training is indicative of both quality and synMR's capabilities as a data distillation technique. SynMR was evaluated both as a whole set and in overlapping subsets containing either a quarter, half, or three quarters of the synMR images.

U-Net: DSC/HD Analysis

The two most popular metrics, DSC and HD, are used to identify ET, WT, and TC segmentation performance for the U-Nets. **Figure 3** shows the DSC and HD for each structure of the validation dataset trained with different subsets of synMR.

Figure 3 shows standard box-plot results for both DSC and HD for each model. *T*-Test reported statistically significance in models containing one quarter, half, and total synMR when compared against baseline. In addition, the relationships between neighboring models as synMR increased showed statistical significance as well. It can be seen that U-Nets trained using at least half synMR show a direct relationship between the amount synMR used and the U-Net performance. The statistical significance in model relationships combined with differences in model performance (**Figure 3**) indicate that a threshold ratio of 2:1 (rMR:synMR) is necessary to introduce more variance while maintaining a proper distribution of data. HD shows significant improvement; however, DSC shows lower relative improvement as DSC is inherently biased in this study due to the fact that it was used as the loss function during the training of each U-Net.

U-Net: Sensitivity/Specificity Analysis

While sensitivity and specificity are not as integral in judging the quality of generated contour as DSC/HD, they show the level of accuracy in defining the tumor border, as well as the size differences between the ground truth and generated contours. As the U-Net was trained, DSC was optimized, however, this metric only indicates the level of volumetric overlap, which leaves the size of the generated contour dependent on other factors. These factors can relate to the differences in training datasets and give insight into how incorporation of synMR changes the U-Nets. **Table 3** shows that when one half or a quarter of synMR was implemented, sensitivity and specificity both increased, indicating improvement in border definition. However, when all synMR was used, sensitivity decreased while specificity remained relatively unchanged. Since synMR showed higher distinction from rMR at the boundaries with sharper gradient drop off, this could lead to a systematic difference of the segmentation labels between the two datasets and led to smaller contours generated from U-Net. However, the smaller contour generated by training on synMR possessing gentler gradients does not negatively affect overall U-Net performance, as specificity and sensitivity mainly show the direction of the offset between the ground truth and generated contours (HD).

Individual Cases

It is necessary to outline the best and worst cases to assess the model performance. The best performing and worst performing

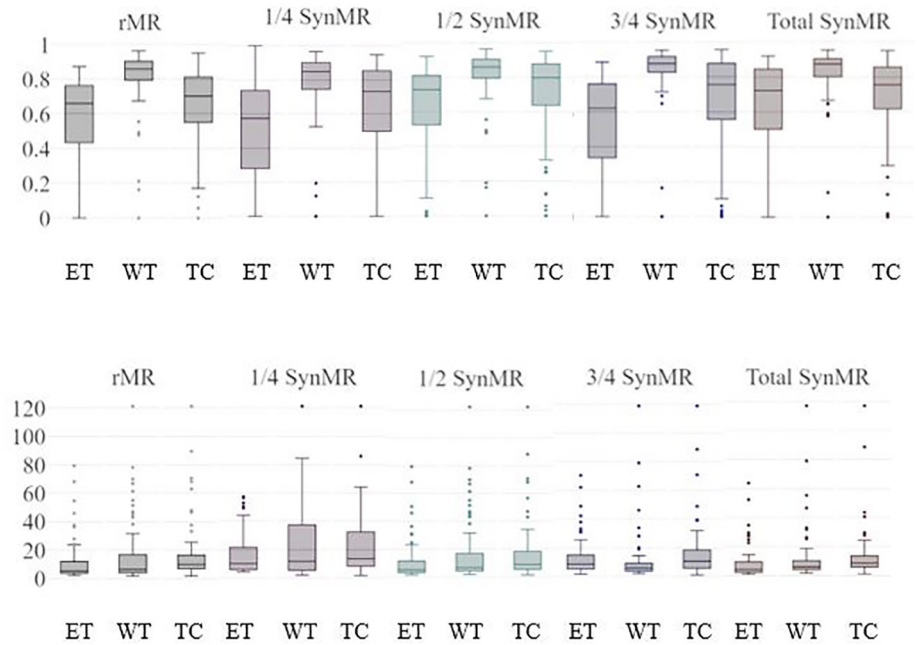


FIGURE 3 | ET/WT/TC Validation Results for U-Nets trained by BraTS MRI and different subsets of synMR(None, 1/4,1/2,3/4, total synMR incorporated). Top row shows DSC, bottom shows HD. Each grouping displays results in order ET/WT/TC. Incorporation of synMR above threshold ratio of 2:1 (rMR:synMR) improves DSC and HD. Standard box-plot format used with singular points displaying outliers.

TABLE 3 | Validation results for U-Nets trained by BraTS MRI and different subsets of synMR.

Percent of SynMR added		None (rMR Only)	1/4	1/2	3/4	All
Sens.	ET	0.70	**0.80**	**0.84**	**0.78**	0.62
	WT	0.89	**0.90**	0.89	0.83	0.80
	TC	0.66	**0.74**	**0.75**	**0.77**	0.66
Spec.	ET	0.96	**0.99**	**0.99**	**0.99**	**0.99**
	WT	**0.99**	0.98	**0.99**	**0.99**	**0.99**
	TC	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**

Bolded values indicate highest contour metric value.

individual cases are carefully evaluated. **Figures 4**–**7** show the comparison of contours on WT for two good and two poor performing cases.

We have observed encouraging improvement of the segmentation accuracy for high grade glioma when the lesion was centrally and radially located. However, the challenge still exists in the low-grade glioma cases due to increased difficulty in boundary definition. Location also plays a role in discerning whether the contouring accuracy would improve or not. The improved low-grade glioma case was centrally located, while in the case that did not show improvement was located toward the edge of the brain. It can also be seen in the improved cases (**Figures 4**, **5**) that the U-Net focuses more on differences in structure, rather than differences in pixel intensity. This is in line with the strength of synMR, as synMR quality regarding structure outperforms its quality pertaining to intensities.

DISCUSSION

The original idea of synthesizing images indistinguishable from reality is inspired by the development of GANs (Goodfellow, 2014). GANs have been employed to expand training datasets for many tasks. Specifically, synthesizing new images as training samples provides a possible solution to overcome the challenge of the limited number of annotated medical images. Frid-Adar et al. (2018) achieved impressive results in lesion classification using GAN-synthesized images, which indicated the potential of GAN for data distillation tasks. Bowles et al. (2018) used GAN for segmentation, however, there are important differences between the studies. First, they experimented on image patches sampled from the dataset, while we experimented on the entire images (Bowles, 2018). Second, their study generated synMR images and contours from Gaussian noise (Bowles, 2018). Due to this, their study was not able to provide a quantitative evaluation between rMR and synMR images. Their work was limited to only providing a visual comparison using the Turing test (section SynMR Qualitative Analysis).

Researchers have leveraged GANs in a conditional setting which allows the model to deterministically control the generation of particular samples based on external information (Gauthier, 2014; Mirza, 2014; Isola, 2017) However, some researchers suggested that adversarial training might be unstable

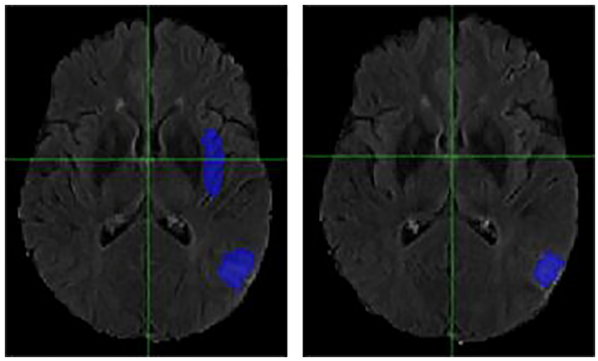


FIGURE 4 | Case One (Improved). Flair MRI. rMR only **(Left)** and Total SynMR MRI **(Right)** DSC of WT was improved from 0.21 to 0.67.

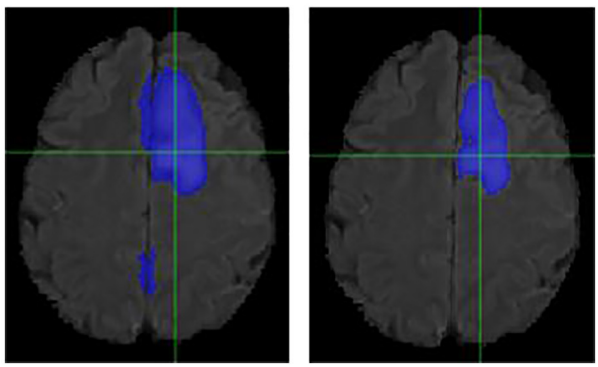


FIGURE 5 | Case Two (Improved). Low-grade Glioma. Flair MRI. rMR only **(Left)** and Total SynMR MRI **(Right)** DSC of WT was improved from 0.49 to 0.88.

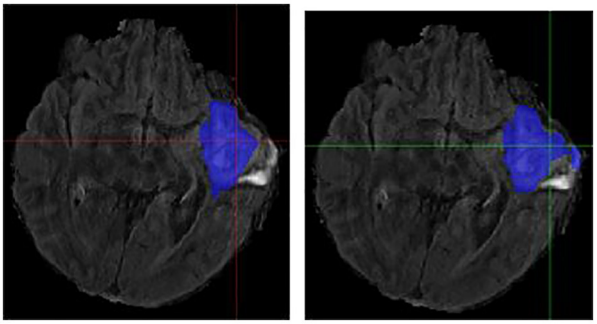


FIGURE 6 | Case One (Worsened). Flair MRI. rMR only **(Left)** and Total SynMR MRI **(Right)** DSC of WT changed from 0.76 to 0.58.

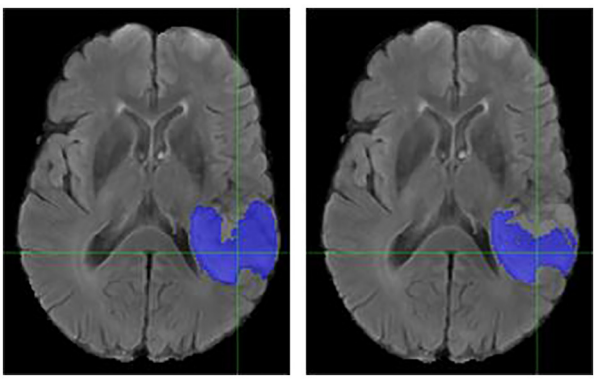


FIGURE 7 | Case Two (Worsened). Low-Grade Glioma. Flair MRI. rMR only **(Left)** and Total SynMR MRI **(Right)** DSC of WT showed a decrease from 0.86 to 0.59.

or even diverge, and introduced image per-pixel loss and perceptual loss (Dosovitskiy and Brox, 2016) that was used in this study. In this paper, we proposed an augmentation network that was trained in a supervised fashion using paired semantic labels and rMR images. This method was chosen over (Frid-Adar, 2018) and (Bowles, 2018) as it incorporated flexible object manipulations with desired scenarios, which allows for the creation of synMR with various tumor size and location from those present in the rMR images.

Our study explored the idea to utilize GAN to generate synthetic images to improve image segmentation performance. Specifically, the segmentation model employed the previously outlined 2D U-Net (Ronneberger et al., 2015) due to its effectiveness, competitiveness and high familiarity, as it has been widely adopted in the field of image segmentation. In the medical field, especially in clinical trials, where limited numbers of training samples are accessible, deep learning models can easily overfit during training and perform poorly in application on independent datasets (Ronneberger et al., 2015). Purely increasing the size of the training dataset by simple inclusion of synMR does not guarantee higher performance. However, the neural networks performance will show improvement if the synMR is of sufficient quality and introduce diversity. This study assumes that the model performance is sufficient to judge the overall data distillation ability of synMR generated in this study. We postulate that that all neural network-based segmentation models should show improvement if trained on datasets containing more variance, although level of improvement may vary model to model. However, this study recognizes that this should be further investigated in a future study by introduction of one or more additional segmentation neural networks.

"TumorGan" (Qingyun Li et al., 2020) and "ANT-GAN" (Sun et al., 2020) were different GAN methodologies to generate synMR. Direct quantitative comparison of synMR image quality is difficult among studies due to synMR/rMR structural differences. Specifically, the tumor location in our synMR was purposefully changed from original rMR to increase variability of resulting datasets. This structural difference between synMR and rMR created the need for advanced qualitative analysis by authoring physicians. However, compared to the other two studies, our work showed competitive results on the improvement of segmentation using synMR as a data augmentation technique. The other two studies reported an increase over baseline of 2.6 and 2.5% in the average DSC while our study showed an improvement of 4.8%.

Statistical investigation of incorporation of certain subsets of synMR during training showed improved performance when incorporating synMR at or above the threshold ratio of 2:1 (rMR to synMR). This ratio could be due to the inherent necessity to introduce additional variance in the training dataset. However, differences in individual synMR image quality could also play a role. Even though the subsets of synMR were staggered in the training models, there are still differences in the quality of individual synMR images. This difference in individual synMR qualities could play a part in the reasoning behind reduced results in segmentation performance when trained using only one randomized subset of synMR. The difference in individual synMR qualities can be partially explained by the fact that rMR quality was not constant. Individual rMR quality differed as it was obtained from different MRI machines over many years. Since image quality had been improved throughout this time, recently obtained rMR images generally show a higher image quality than older rMR. Performance of the employed GAN was impacted by this as it is not likely that the generated synMR will possess greater quality than its corresponding input rMR. However, the relationship of individual synMR quality and its impact regarding the model's performance should be further investigated.

CONCLUSION

We were able to generate high quality Flair, T2, T1, and T1CE synMR using the presented augmentation network and had a thorough evaluation of the images both quantitatively and qualitatively. In addition, the synMR images proved their capability as a data augmentation technique, as incorporation of the created synMR images to increase the size and diversity of the training dataset showed promising results. The presented data manipulation strategy has the potential to address the challenges regarding the limited labeled medical dataset availability for medical image segmentation.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

NW designed the research and methodology. EC and ZD performed the data pre-processing, machine learning training, and analysis. EC, NW, ZD, and JS wrote the paper. EL and JS provided clinical guidance on the evaluation of image quality and segmentation accuracy.

FUNDING

This work was supported by a Research Scholar Grant, RSG-15-137-01-CCE from the American Cancer Society.

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiaomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A, Bauer, S., Rempfler, M., Crimi, A., et al. (2018). 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.*

Bakas, S., Sotiras, A. H., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., et al. (2017a). "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection," in *The Cancer Imaging Archive*.

Bowles, C. (2018). GAN augmentation: augmenting training data using generative adversarial networks. *arXiv [Preprint]. arXiv:1810.10863v1.*

Bowles, C., Pereanez, M., Bowles, C., Piechnik, S. K., Neubauer, S., and Petersen, S. E. (2018). GAN augmentation: augmenting training data using generative adversarial networks. *arXiv [Preprint]. arxiv: 1810.10863.*

Dosovitskiy, A., and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. *arXiv [Preprint]. arXiv:1602.02644v2.*

Frid-Adar, M. (2018). GAN-Based synthetic medical image augmentation for increased CNN performance in liver lesion classifcation. *Neurocomputing* 321, 321–331.

Frid-Adar, M., Diamant, I., Klang, L., Amitai, M., Goldberger, J., and Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classifcation. *Neurocomputing* 321, 321–331. doi: 10.1016/j.neucom.2018.09.013

Gauthier, J. (2014). *Conditional Generative Adversarial Nets for Convolutional Face Generation.* Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition.

Gomez-Iturriaga, A., Casquero, F., Urresola, A., Ezquerro, A., Lopez, J. I., Espinosa, J. M., et al. (2016). Dose escalation to dominant intraprostatic lesions with MRI-transrectal ultrasound fusion High-Dose-Rate prostate brachytherapy. Prospective phase II trial. *Radiother. Oncol.* 119, 91–96. doi: 10.1016/j.radonc.2016.02.004

Goodfellow, I. (2014). Generative adversarial nets. *arXiv [Preprint]. arXiv:1406.2661v1.*

Isola, P. (2017). "Image-to-image translation with conditional adversarial networks" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5967–5976.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision* (Amsterdam: Springer), 694–711.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning*.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Mirza, M. (2014). *Conditional Generative Adversarial Nets.* CoRR abs/1411.1784.

Ostrom, Q. T., Gittleman, H., Truitt, G., Boscia, A, Kruchko, C., and Barnholtz-Sloan, J. S. (2018). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2011-2015. *Neuro Oncol.* 20, iv1–iv86. doi: 10.1093/neuonc/noy131

Pelt, D. M., and Sethian, J. A. (2018). A mixed-scale dense convolutional neural network for image analysis. *Proc. Natl. Acad. Sci. U.S.A.* 115, 254–259. doi: 10.1073/pnas.1715832114

Qingyun Li, Z. Y., Yubo, W., and Haiyong, Z. (2020). TumorGAN: a multi-modal data augmentation framework for brain tumor segmentation. *Sensors.* 20:4203. doi: 10.3390/s20154203

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. *arXiv [Preprint]. arXiv:1505.04597v1*.

Shen, D., Wu, G., and, Suk, B. E. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv [Preprint]. arXiv: 1409.1556v6*.

Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., and Paisley, J. (2020). An adversarial learning approach to medical image synthesis for lesion detection. *IEEE J. Biomed. Health Inform.* 24, 2303–2314.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership