Hirni: Segmentation of Brain Tumors in Multi-parametric Magnetic Resonance Imaging Scans

Gabriel Mejía *, Danniel Moreno †, Daniela Ruiz ‡, Nicolás Aparicio §
Departamento de Ingeniería Biomédica, Universidad de los Andes. Bogotá, Colombia
Email: *gm.mejia@uniandes.edu.co, †dr.moreno@uniandes.edu.co, ‡da.ruizl1@uniandes.edu.co, §n.aparicioc@uniandes.edu.co

***Abstract*—Glioma is a type of tumor that develops in the brain and causes cancer in nervous system. They are normally associated with a bad prognosis due to its heterogeneity and the difficulty to resect them. Taking this into account, we developed an algorithm capable of segmenting the brain into four parts: background (brain without tumor), necrotic core (NRC), edematous region (ED) and Active and non-enhancing/necrotic tumor regions (AT). First we approached the problem as a binary case (tumor/non-tumor segmentation) and then we used these masks to improve our multi-class case. For descriptors of each pixel, we used the intensities associated to the four different MRIs contrast media and the prior probability distribution of the training set (two for binary and four for multi-class), and we trained a Random Forest classifier with them. Taking into account the limitation of the computational resources with which we work, our results are outstanding (Dice coefficient 0.806 in the binary case) and competitive with state of the art methods.**

Keywords: Glioma, MRI, Semantic Segmentation, Brain Tumor Segmentation Challenge, Random Forest.

I. Introduction

Gliomas are the leading cause of cancer in the central nervous system [1]. Of the approximately 13,000 new cases of brain tumors in the United States each year, about 65% are of glial origin. Glioblastoma (GB) is the most aggressive type; it is a fast-growing and potentially lethal form of brain cancer that arises from astrocytes and invades normal brain tissue. As the tumor grows, it can infiltrate, compress, and destroy the healthy brain tissue surrounding it. Moreover, these invading tumor cells can irritate the brain causing seizures. Approximately 20%–30% of patients with brain neoplasms experience epileptic-like seizures [2].

Additionally, epidemiological studies show a high presence of gliomas in the Latin American population. Of the 62.1% of nervous system tumors, approximately 89% correspond to gliomas. Considering the age group, 13% of tumors occur in children (0-14 years); 58% in the 15-64 age group, and 27% in the elderly (more than 65 years). GB is normally associated with a bad prognosis, only 33% of patients survive in a year and only 5% of patients live for more than 5 years after diagnosis [3]. Besides, It has been observed that this survival rate can be even less than 5% in Latin America [4]. Accordingly to the above, the survival of patients depends to a large extent on the correct and timely diagnosis of the brain tumor.

Glioblastomas are composed of cells of great morphological variability, elevated mitotic activity, microvascular proliferation, severe and characteristic endothelial hyperplasia, intravascular microthrombi, and extensive ischemic or pseudoempalized necrosis [3]. Due to the heterogeneity of glioblastomas, it is hard to diagnose a patient correctly in an early stage of the cancer. However, it is possible to evidence some tumor's characteristics in the sub-regions representation in varying intensity profiles disseminated across multi-parametric magnetic resonance imaging (mpMRI) scans, providing more information to perform the task of tumor segmentation using an algorithm.

In fact, There is a large amount of literature on computational algorithms addressing the challenging task of brain tumor segmentation in MRI. The two distinguishable milestones, that have transformed the way this algorithms are evaluated and how the segmentation problem is approached, are the creation of the BraTS challenge and dataset, and the extensive implementation of Deep Learning techniques like Convolutional Neuronal Networks (CNN) as an effective segmentation method during the recent years.

Previous to these milestones, statistical methods comprised of classification and clustering algorithms were used. Classification requires training data to learn a model based on which new instances can be labeled, while clustering groups data based on certain similarity criteria. Most of the glioma segmentation algorithms are based on these techniques because multimodal datasets can be handled easily. The general idea is to decide for every single voxel to which class it belongs based on its feature vector [5]. In 1993, Shad *et al.* [5] analyzed texture patterns of different brain tissues. Phillips *et al.* (1995) [5] employed fuzzy c-means clustering and Vaidyanathan (1995) [5] compared that method with kNN-clustering for tumor volume determination on multispectral 2D image slices. During the following years, this approach was combined with knowledge-based techniques, and between 2007 and 2012 [5] developed methods implementing Support Vector Machines (SVM), decision forest classification and neural networks.

However, taking into account the results of the BraTS 2018 challenge and its previous editions, it is evident that the most successful approach to the brain tumor segmentation

978-1-6654-0855-4/21/$31.00 ©2021 IEEE

problem during the recent years has been building an adequate architecture of neuronal networks, where the implementation of CNN and Fully Convolutional Networks (FCN) has played an important role. Furthermore, in the review article of BraTS 2018 it is noted that while the accuracy of individual automated segmentation methods has improved, their level of robustness is still inferior to expert performance (inter-rater agreement) [6]. The current state-of-the-art, has a dice coefficient of 0.766 to 0.883 [7] (depending on the region of the tumor). However, the fusion of segmentation labels from various individual automated methods shows robustness superior to the ground truth inter-rated agreement provided by clinical experts, in terms of both accuracy and consistency across subjects [6].

II. Materials and Methods

A. Dataset prepossessing

Our method Hirni, which stands for brain in ancient germanic language, was trained and evaluated using the Brain Tumor Segmentation (BraTS) Challenge 2020 dataset available at the Center for Biomedical Image Computing Analytics (CBICA) Image Processing Portal of the University of Pennsylvania [6] [8] [9]. The dataset with available annotations was subdivided in training, validation, and test sets, which consist of about the 60%, 21% and 19% of the total patient information (Tab. I). The information of every patient consists of ground truth annotations of 3 sub-regions of brain tumors and 4 MRIs with different contrast agents: a native T1-weighted scan (**T1**), a post-contrast T1-weighted scan (**T1Gd**), a native T2-weighted scan (**T2**), and a T2 Fluid Attenuated Inversion Recovery (**T2-FLAIR**) scan [6]. Each of these MRIs is comprised of 155 consecutive 240 × 240 pixel cross-sectional images. The ground truth annotations were made by experts and manually-revised by board-certified neuroradiologists [6]. These annotations consist of the segmentation masks of three different sub-regions of the tumors: whole tumor extent (**WT**), gross tumor core outline (**TC**), and active and non-enhancing/necrotic tumor regions (**AT**) [6].

TABLE I
Data sets distribution.

	Percentage	Patients
Training	60%	220
Validation	21%	72
Test	19%	70

B. Data normalization and balancing

Considering that the MRI intensity values vary due to differences among the acquisition methods, an intensity normalization was done over each MRI, subtracting the mean and dividing by the standard deviation of the whole brain region each of the voxel intensities. Taking into account that the number of voxels of the regions of interest (glioma tissue) is much lower in comparison to the voxels corresponding to the healthy tissue (background), a sampling parameter was introduced. During sampling, randomly distributed voxels of the background were taken to match the number of voxels of the least represented class of interest in every MRI. This was done to avoid bias during classification and diminishing the need of computational resources.

C. Descriptors

Descriptors of the sampled and normalized voxels of each MRI were calculated for 3 different representation spaces: a four-dimensional space corresponding to the intensities of the voxels using the different contrast agents (T1, T1Gd, T2, and T2-Flair), a space corresponding to the probabilities of belonging to each class based on the frequency distributions of the intensities in the original training set (Prior), and a space consisting of the concatenation of the coordinates of the two previous spaces. We approached the classification as a binary (tumor/non-tumor) problem, and a multi-class problem (3 tumor regions of interest and healthy tissue). Thus, the dimension size of the prior representation space varies among the two different approaches to the classification problem. Fig. 1 shows the structure of the descriptor that concatenates the intensity of the voxel with the different contrast agents (which can be interpreted as a color space), and the prior distribution for the multi-class problem. Fig. 2 shows the joint histograms associated with the prior distributions, which were built using only the three contrast agents that provide more differentiation.

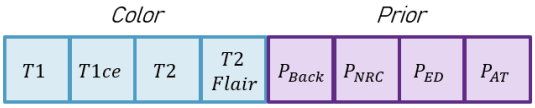

Fig. 1. Descriptor containing both the color intensity information for each voxel channel (T1, T1Gd, T2, and T2-Falir) and the probability that the voxel belongs to each of the classes (NRC, ED, AT, and Background).

D. Training and post-processing

In order to classify each voxel into a segmentation class based on its descriptor, we used a Random Forest classifier. Different models of this random forest classifier were trained varying the number of trees, the maximum depth of the tree and the maximum number of features to consider when looking for the best divergence. To generate an inference corresponding to the tumor segmentation of a new MRI, the descriptors of each voxel of the MRI are calculated and then passed through the pre-trained Random Forest. Binary and multiclass classification random forests were trained parallelly out an Intel Core i7-8750H CPU. The segmentation masks obtained through binary classification were post-processed applying a morphological opening and small connected-components removal, taking into account that tumor regions are comprised of numerous adjacent voxels. The maximum size of the connected components removed was another parameter that was varied during experimentation. Binary classification showed an outstanding recall, so an intersection between the post-processed binary segmentation mask and the multiclass

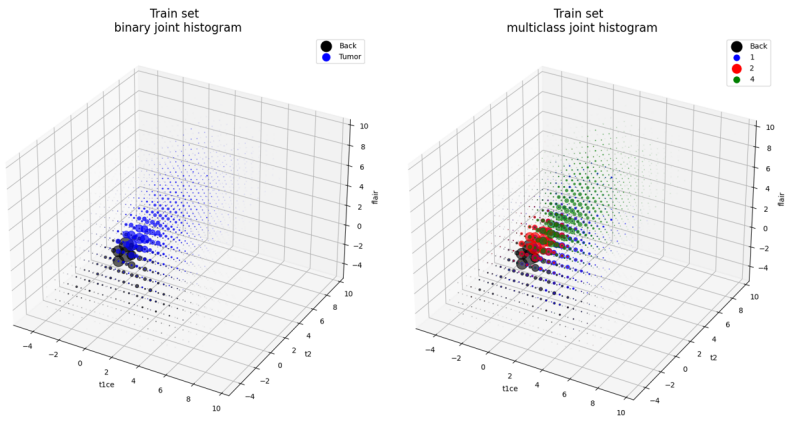

Fig. 2. Joint histograms showing the prior distributions for the case of **Left:** Binary segmentation. In blue is the probability distribution of the voxels for the tumor class, in black is the probability distribution for the non-tumor class. **Right:** Multi-class segmentation. In black is the probability distribution of each voxel belonging to the background, in blue the probability of each voxel for the NRC class, in red for the ED class, and in green the probabilities of belonging to the AT class.

segmentation mask was made to guarantee that the different predicted tumor regions lay inside the predicted whole tumor mask, improving the model precision.

E. Evaluation

The performance of output annotations was assessed by 4 complementary metrics: Dice coefficient, precision, recall, and Jaccard index. A complete confusion matrix for the train and validation set was constructed and used to find global values of the metrics for each parameter combination. The model optimization procedure used the binary case Dice coefficient as the maximization function.

III. Results

A. Experiments over validation set

The performance of four different models applying variation of parameters was evaluated.

In the first case, the *fraction* parameter was changed: 50% of the sampled voxels were taken to perform the training and the validation. The metrics for this case are reported in Tab. II.

TABLE II
EXPERIMENT PERFORMED IN THE VALIDATION SET. THE PARAMETERS USED ARE; *fraction*: 0.5, *max_depth*: 20, *n_estimators*: 50, AND *max_features*: 4

	Dice Coefficient	Precision	Recall	Jaccard Index
WT	0.621	0.489	0.851	0.451
NCR	0.304	0.501	0.218	0.179
ED	0.455	0.326	0.751	0.294
AT	0.571	0.437	0.820	0.399

A second evaluation was performed with a model that was trained with 70% of the sampled voxels. The metrics for this case are reported in Tab. III

According to the two experiments mentioned above, it is decided that a fraction value of 0.5 presents better metrics than 0.7. In a third evaluation, the number of trees in the Random Forest was changed. The two previous experiments had a value for this parameter of 50. In this experiment a value of 24 is assigned, as observed in Tab. IV

TABLE III
EXPERIMENT PERFORMED IN THE VALIDATION SET. THE PARAMETERS USED ARE; *fraction*: 0.7, *max_depth*: 20, *n_estimators*: 50, AND *max_features*: 4

	Dice Coefficient	Precision	Recall	Jaccard Index
WT	0.621	0.488	0.854	0.450
NCR	0.306	0.504	0.220	0.181
ED	0.452	0.324	0.753	0.293
AT	0.570	0.437	0.820	0.399

TABLE IV
EXPERIMENT PERFORMED IN THE VALIDATION SET. THE PARAMETERS USED ARE; *fraction*: 0.5, *max_depth*: 20, *n_estimators*: 24, AND *max_features*: 4

	Dice Coefficient	Precision	Recall	Jaccard Index
WT	0.623	0.492	0.849	0.453
NCR	0.302	0.499	0.217	0.178
ED	0.455	0.327	0.767	0.294
AT	0.584	0.455	0.817	0.412

The fourth experiment maintains a number of trees of 24, a fraction value of 0.5, and a maximum depth of 20. However, we changed the value of maximum features from 4 to 2. The metrics for this experiment are shown in Tab. V. It is important to mention that these four experiments used a composite descriptor, i.e., it takes into account both color intensity and probability information for each voxel.

Finally, the best model was chosen with the best Dice coefficient in the binary case. According to the above, it can be observed that the third experiment (observe Tab. IV) presents a higher value for that case. Therefore, this will correspond to our best model to use in the test set.

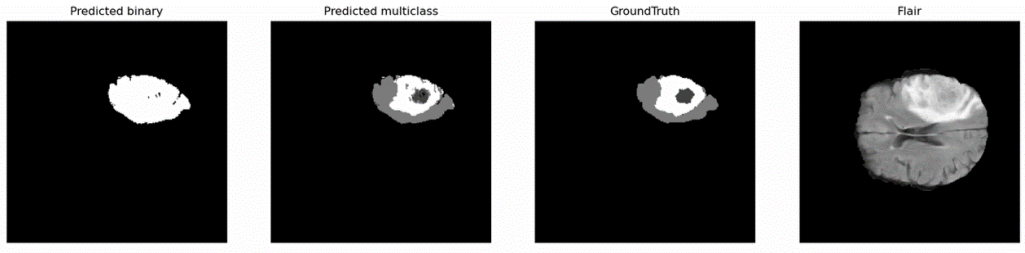

Fig. 3. From left to right: binary segmentation achieved with Hirni, multi-class segmentation performed by Hirni, Ground truth annotation, corresponding cross section image in T2-Flair channel.

TABLE V
EXPERIMENT PERFORMED IN THE VALIDATION SET. THE PARAMETERS USED ARE; *fraction*: 0.5, *max_depth*: 20, *n_estimators*: 24, AND *max_features*: 2

	Dice Coefficient	Precision	Recall	Jaccard Index
WT	0.620	0.487	0.851	0.450
NCR	0.306	0.509	0.219	0.180
ED	0.454	0.325	0.753	0.294
AT	0.572	0.439	0.820	0.400

B. Evaluation experiments

Taking into account the best model of the validation set, which contains a good combination of parameters, we proceeded to modify the model, changing the removal of related component from 100.000 to 10.000, and we developed an opening with a disk with radius 2. The performance metrics are as follows:

TABLE VI
EXPERIMENT PERFORMED IN THE TEST SET. THE PARAMETERS USED CORRESPOND TO THE BEST MODEL IV WITH CHANGES IN RELATED COMPONENTS AND THE IMPLEMENTATION OF THE OPENING.

	Dice Coefficient	Precision	Recall	Jaccard Index
WT	0.806	0.857	0.762	0.676
NCR	0.321	0.539	0.228	0.191
ED	0.620	0.611	0.629	0.449
AT	0.704	0.660	0.755	0.544

Fig. 3 shows an example of the results obtained by the model with the best performance.

IV. CONCLUSION

In the first place, it can be mentioned that the fraction parameter allows being more flexible when processing data since this parameter indicates a percentage of the amount of data to be taken to perform the training. Regarding the variation of its value, it is possible to conclude that similar metrics are obtained between a fraction value of 0.5 and 0.7. However, the best metrics are obtained with a value of 0.5, which can be explained by a possible over-fit in the model. Additionally, increasing the value of fraction does not show a big change in the metrics but it does in terms of time and computational cost ($\approx$ 2 hours of computation time).

Regarding the comparison with the related work, our results for every class cannot be directly compared to Brats2018 metrics due to the fact that regions of interest differ. However, our binary case (WT) evaluation is equivalent to the challenge method. With that in mind, the best version of Hirni (Dice of 0.806) shows a competitive but inferior performance to that of Brats2018 winning method (Dice of 0.910). It is important to clarify that the database used in our case is a subset of the Brats 2018 challenge database. Additionally, for the development of our algorithm, we took a percentage of the full training set to create the validation set.

On the other hand, we have found that the main restriction of Hirni corresponds to the limited computational resources that we had available. Due to appearance of the 3D images, it is evident that using a descriptor with local information like texture or neighbors intensities may improve the results of our method, but considering that we are generating a descriptor for each voxel and working with a large data set, increasing the size of descriptors is not computationally viable. Further iterations employing more computational resources could optimize our metrics by grid search or other hyper-parameter optimization method.

REFERENCES

[1] R. Rodríguez, K. Lombardo, G. Roldán, J. Silvera, and R. Lagomarsino, "Glioblastoma multiforme cerebral hemisférico: análisis de sobrevida de 65 casos tratados en el Departamento de Oncología del Hospital de Clínicas, desde 1980 a 2000," *Revista Médica del Uruguay*, vol. 28, no. 4, pp. 250–261, 2012.

[2] E. P. dkk Widmaeir, *Vander'S Human Physiology: the Mechanisms of Body Function, Fifteenth Edition*, vol. 53. 2017.

[3] C. C., "Glioblastoma : Análisis Molecular Y Glioblastoma : Molecular Analysis and Its Clinical," *Rev Peru Med Exp Salud Publica*, vol. 32, no. 2, pp. 316–325, 2015.

[4] M. I. Ocampo Navia, J. C. Gómez Vega, and O. H. Feo Lee, "Epidemiología y caracterización general de los tumores cerebrales primarios en el adulto," *Universitas Médica*, vol. 60, no. 1, 2018.

[5] S. Bauer, R. Wiest, L. P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," jul 2013.

[6] B. et al, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," *arXiv*, 2018.

[7] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11384 LNCS, pp. 311–320, oct 2018.

[8] B. H. Menze, A. Jakab, S. Bauer, and e. a. Kalpathy-Cramer, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, oct 2015.

[9] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, sep 2017.