Exploring Radiologic Criteria for Glioma Grade Classification on the BraTS Dataset

P. Dequidt, Pascal Bourdon, B. Tremblais, C. Guillevin, B. Gianelli, C. Boutet, J.-P. Cottier, J.-N. Vallée, C. Fernandez-Maloigne, R. Guillevin

Version of Record: https://www.sciencedirect.com/science/article/pii/S195903182100049X
Manuscript_acca2c95c53751cbef14a0e90d8a485e

Exploring radiologic criteria for glioma grade classification on the BraTS dataset

Paul Dequidt[1,3,4,*], Pascal Bourdon[1,4], Benoit Tremblais[1,4], Carole Guillevin[2,4], Benoit Gianelli[2,4], Claire Boutet[5], Jean-Philippe Cottier[6], Jean-Noël Vallée[7], Christine Fernandez-Maloigne[1,4], Rémy Guillevin[2,4]

Abstract

1) Objectives

Glioma grading using maching learning on magnetic resonance data is a growing topic. According to the World Health Organization (WHO), the classification of glioma discriminates between low grade gliomas (LGG), grades I, II ; and high grade gliomas (HGG), grades III, IV, leading to major issues in oncology for therapeutic management of patients. A well-known dataset for machine-based grade prediction is the MICCAI Brain Tumor Segmentation (BraTS) dataset. However this dataset is not divided into WHO-defined LGG and HGG, since it combines grades I, II and III as "lower grades gliomas", while its HGG category only presents grade IV glioblastoma multiform. In this paper we want to train a binary grade classifier and investigate the consistency of the original BraTS labels with radiologic criteria using machine-aided predictions.

2) Material and methods

Using WHO-based radiomic features, we trained a SVM classifier on the BraTS dataset, and used the prediction score histogram to investigate the behavior of

*XLIM Laboratory, 11 Bd Marie et Pierre Curie, 86360 Chasseneuil-du-Poitou, France
Email address: paul.dequidt@univ-poitiers.fr (Paul Dequidt)
[1]XLIM Laboratory, UMR CNRS 7252, University of Poitiers, Poitiers, France
[2]DACTIM-MIS, LMA, UMR CNRS 7348, University of Poitiers, Poitiers, France
[3]Siemens Healthcare SAS, Saint-Denis, France
[4]Common Laboratory CNRS-Siemens I3M
[5]Department of Radiology, University Hospital of Saint Etienne, Saint Etienne, France
[6]Department of Radiology, CHRU de Tours, UMR 1253 iBrain, Inserm, University of Tours, Tours, France
[7]CHU Amiens Picardie, University of Picardie Jules Verne (UPJV), Amiens, France

our classifier on the lower grade population. We also asked 5 expert radiologists to annotate BraTS images between low (as opposed to lower) grade and high grade glioma classes, resulting in a new groundtruth.

3) Results

Our first training reached 84.1% accuracy. The prediction score histogram allows us to identify the radiologically high grade patients among the original lower grade population of the BraTS dataset. Training another SVM on our new radiologically WHO-aligned groundtruth shows robust performances despite important class imbalance, reaching 82.4% accuracy.

4) Conclusion

Our results highlight the coherence of radiologic criteria for low grade versus high grade classification under WHO terms. We also show how the histogram of prediction scores and crossed prediction scores can be used as tools for data exploration and performance evaluation. Therefore, we propose to use our radiological groundtruth for future developpement on binary glioma grading.

Keywords: Glioma grading, machine learning, automatic classification, prediction score, virtual biopsy, radiomics

---

1. Introduction

Gliomas are an aggressive type of brain tumor. In their most advanced form, they are linked to a high death rate within a short survival range. The World Health Organisation (WHO) uses histopathology and genomic criteria to identify the malignancy of the tumor through 4 grades, ranging from grade I to grade IV [1]. Grades I and II are labelled as Low Grade Gliomas (LGG_{WHO}) while grades III and IV are labelled as High Grade Gliomas (HGG_{WHO}). Due to their slow or asymptomatic developement, LGG_{WHO} are often less diagnosed than HGG_{WHO} [2]. The gold standard to assess the grade is biopsy, an invasive technique. Biopsy classifies tumors based on microscopic similarities of the cells and their levels of differentiation. Biopsy is subject to sampling error and inter-observer variation [3]. The histological information is combined with genotype

analysis to screen mutations such as IDH type and the 1p/19q codeletion status [4].

Non-invasive methods based on magnetic resonance imaging (MRI) are being developped to create a robust alternative. Machine learning has been used on MR data to discriminate between LGG_{WHO} and HGG_{WHO}, for example with simple classifiers like SVM [5] or Random Forests [6]. Deep learning with convolutional neural networks (CNN) has also been tested on anatomical data for grade classification [7]. While classifying all four glioma grades have been tested [8], most publications focus on a binary grade discrimination between LGG_{WHO} and HGG_{WHO} [9]. This specific discrimination has important clinical impact for the patient, as the evolution from LGG_{WHO} to HGG_{WHO} is linked to a short survival range [10].

These approches rely on the radiomic analysis of glioma used for tumor segmentation : the tumor shape and heterogeneity, the length of the first and second major axis, or the presence of necrosis and enhancement in T1 contrast-enhanced (T1ce) sequences can be used in the MR diagnosis [11].

One of the most popular datasets used for automatic binary grade classification is the BraTS dataset [12, 13, 14]. This dataset, originally published for a tumor segmentation challenge, has been used extensively for binary glioma grade classification. But it is not providing a WHO-aligned division, as it groups together grades I, II and III under the term "lower grades gliomas" (LGG_{BraTS}), and has only glioblastoma multiform, or grades IV, in its high grades gliomas category (HGG_{BraTS}). Therefore, we want to investigate the consistency of the LGG_{BraTS} population, using machine learning and WHO-aligned radiomic features.

In this paper, we present the current state of the art for radiomics analysis and glioma grading using artificial intelligence. We train a SVM classifier with radiomic features on binary grade discrimination. Then we use prediction scores to analyze and identify radiologically high grade patients among the "lower grade" population of the BraTS dataset. We also developped a visual module to allow radiologist experts to discriminate between LGG_{WHO} and HGG_{WHO}

and collect their votes. We used those votes to align the BraTS dataset on the WHO classification and create a radiological groundtruth. This section also shows the performances of our classifier on this new groundtruth data. A discussion about the results is given in conclusion.

2. State of the art

2.1. Radiomic analysis of anatomical MR images

Perfusion imaging and MR spectroscopy give important information for glioma grading and are needed for a more complete body of evidence regarding the diagnosis [15]. Nevertheless, anatomical MR imaging gives access to some features used for glioma grading. It has been known for a long time that mass effect, cyst formation and necrosis are statistically significant predictors of high malignancy [16]. Inversion Recovery sequences allow the signal of a specific tissue to be cancelled such as in the FLuid Attenuated Inversion Recovery (FLAIR) sequence, where the Cerebro Spinal Fluid signal is cancelled. This makes the FLAIR sequence the main sequence used for lesion and oedema detection. Contrast-Enhanced sequences such as T1ce show neoangiogenesis, which is a marker for high grade, even though up to one third of high grade gliomas do not show enhancement signal. [11]. After the WHO 2016 reference publication, some studies have linked the MR phenotype on anatomical imaging with the genotype. Examples include the sharpness of tumor borders or T2-FLAIR mismatch sign as features for 1p/19q codeletion or IDH mutation [17][18]. Therefore, anatomical MRI can approach the WHO classification of gliomas.

2.2. Glioma grading using machine learning

Machine and deep learning can be used for glioma grading on MR data. A common pipeline for machine learning includes feature extraction and ranking through a feature selection algorithm, such as SVM-Recursive Feature Elimination [5] and gives 95.5% accuracy in the best case [19]. Sun et al. compare

16 feature selection algorithms and 15 different classifiers. They get their best result with the SVM feature selection [20]. As an intelligible classifier, SVM is also used often and give good results, up to 94.8% on anatomical imaging [21].

When relying on anatomical sequences only, texture analysis can be used in a machine learning scheme for a binary grade discrimination [22]. Convolutional neural networks like VGG-16 give results up to 95% accuracy [23] and random forests reach 88.77% [6]. As such, convolutional neural networks reach interesting performances for glioma grading, but are still computationaly heavy and lack intelligibility. This is why we chose to use a machine learning classifier like SVM, as it produces more intelligible results with less computing power requirements.

2.3. Glioma grading learning dataset

We used the 2018 version of the BraTS dataset, composed of 285 glioma cases. These patients are divided into 210 glioblastoma multiform (HGG_{BraTS}, grade IV) which is the most advanced grade for gliomas, and 75 "lower grade glioma" (LGG_{BraTS}, grades I, II and III). These labels have been established by histological screening. For each patient, 4 registered and skull-stripped anatomical sequences are available : T1, T1ce, T2 and T2 FLAIR.

As "lower grade gliomas" and "low grade glioma" both share the same acronym, some authors have trained glioma grading classifiers on the BraTS division, while stating their work was a WHO-based classification [24][25]. Therefore, there's also a need to clarify this distinction. Table 1 shows how the grades are grouped under the WHO and BraTS groundtruth data. We can see that the LGG_{BraTS} population is a mixed population, with LGG_{WHO} and HGG_{WHO} patients and we don't have access to the precise grade of each patient. In order to explore the consistency of each label, we are going to analyze how a SVM classifier sorts each patient when given WHO-based radiomics criteria.

Table 1: Glioma grades among the WHO classification and the BraTS groundtruth data

Grade I	Grade II	Grade III	Grade IV
LGG_{WHO}		HGG_{WHO}	
LGG_{BraTS}			HGG_{BraTS}

3. Computer-aided low vs high grade binary classification

In this section we investigate the consistency of the original BraTS labels with radiologic criteria. Using WHO-aligned radiomic features, we want to explore how the LGG_{BraTS} population is processed by a SVM classifier. In order to do so, we train a SVM classifier and evaluate its performance by analyzing its prediction score on the LGG_{BraTS} population. The prediction score histogram shows HGG_{WHO} patients within the false positives. This result highlights the coherence of radiologic criteria used for binary classification under WHO terms.

3.1. Data

Our first training was performed on the BraTS division. This division holds a partial truth of the final grade division we want to achieve. Every patients in the HGG_{BraTS} being glioblastoma multiform, which are the ultimate evolution of gliomas, this group is highly representative of high gradeness under WHO terms, while the LGG_{BraTS} population is a mixed population. Therefore, we propose to train a classifier and analyze its prediction score performances on the LGG_{BraTS} population and see how it dealt with the unmatching HGG_{WHO} patients within it.

3.2. Features

Using the segmentation groundtruth data given with the BraTS dataset, we computed 51 features from the PyRadiomics package [26] for each patient : 7 shape features, 6 histogram-based intensity features and 5 texture features. The shape features are the length of major and minor axis, the maximum 3D diameter, elongation, flatness, sphericity and surface area. These 7 features are the

same for all 4 sequences available and are only computed once per patient. The remaining features are computed on each 4 sequences. The histogram-based intensity features include mean, skewness, kurtosis, contrast, energy and entropy. For texture analysis we used the correlation of the gray level co-occurrence matrix (GLCM), coarseness, inverse difference moment (IDM), complexity and strength. This set of features (shape, intensity, texture) relates to the criteria used by radiologists during glioma grade assessment and have been selected as intelligible features. For example, IDM and GLCM correlation are homogeneity markers while complexity and strength give primitive-based information. Choosing these features allows us to model the radiologic analysis under WHO-terms, as necrosis and gadolinium enhancement patterns make $\mathrm{HGG_{WHO}}$ more heterogenous lesions than $\mathrm{LGG_{WHO}}$ lesions.

3.3. Results and consistencies on radiomics criteria

We trained an SVM classifier and analyzed its performances through its prediction score results. We tested a C-Support Vector Classifier with different hyperparameters : the used kernel (linear or Radial Basis Function) and the value of the regularization parameter C (between 0.1 and 2). The hyperparameters were selected and optimized with a 5-fold cross-validation and a grid search. We applied usual techniques to avoid class imbalance and overfitting, namely the use of balanced class weights in training and the 5-fold cross validation. The best result was obtained with linear kernel and a C value of 1.0. We reached 84.1% accuracy, 87.0% sensitivity and 75.9% specificity on the BraTS dataset. With this first classifier, we can explore these results on the $\mathrm{LGG_{BraTS}}$ population to analyze how it dealt with its unmatching $\mathrm{HGG_{WHO}}$ patients.

For our test, we define true positives and true negatives as shown in Table 2. We define the test as positive when the patient is labelled as $\mathrm{HGG_{BraTS}}$ by our classifier. We want to discriminate the $\mathrm{LGG_{WHO}}$ and $\mathrm{HGG_{WHO}}$ population among the BraTS dataset. As some $\mathrm{HGG_{WHO}}$ are overlapping into the $\mathrm{LGG_{BraTS}}$ group, we must focus on this population. With our classifier, this population is divided between true negatives and false positives. False pos-

Table 2: Confusion matrix terminology

		BraTS groundtruth data	
		LGG_{BraTS}	HGG_{BraTS}
SVM	LGGsvm	True negatives	False negatives
	HGGsvm	False positives	True positives

Table 3: Confusion matrix for the SVM classifier trained on the BraTS dataset

		BraTS groundtruth data	
		LGG_{BraTS}	HGG_{BraTS}
SVM	LGGsvm	54	11
	HGGsvm	21	199

itive patients are particularly interesting, as they are LGG_{BraTS} classified as HGG_{BraTS} by our classifier. The confusion matrix of our first training is shown in Table 3.

In order to analyze our classifier behavior, we propose to use the prediction score probabilities, as defined by Platt et al. [27]. This score gives a value between 0 and 1 and can be read as the confidence of our classifier when assigning a patient to a class. In order to study the prediction scores of the whole LGG_{BraTS} population, we plotted the histogram of prediction for every LGG_{BraTS} patient (Fig. 1a). This histogram gives us information about the quality of discrimination given by our classifier. Indeed, the closer to 0 or 1 a patient, the more certain our classifier is when assigning a patient to a group. A patient closer to 0.5 indicates that our classifier is less confident about its assignement. We can analyze the shape of this histogram as a bimodal distribution with one mode in the true negatives part (correctly labelled LGG_{BraTS}) and another in the false positives part in red (LGG_{BraTS} labelled as HGG_{BraTS}). To get a better understanding of the behavior of our classifier, we can analyze the patients on the far-left side of the histogram : these patients were predicted as positive by our classifier with a very high confidence level.

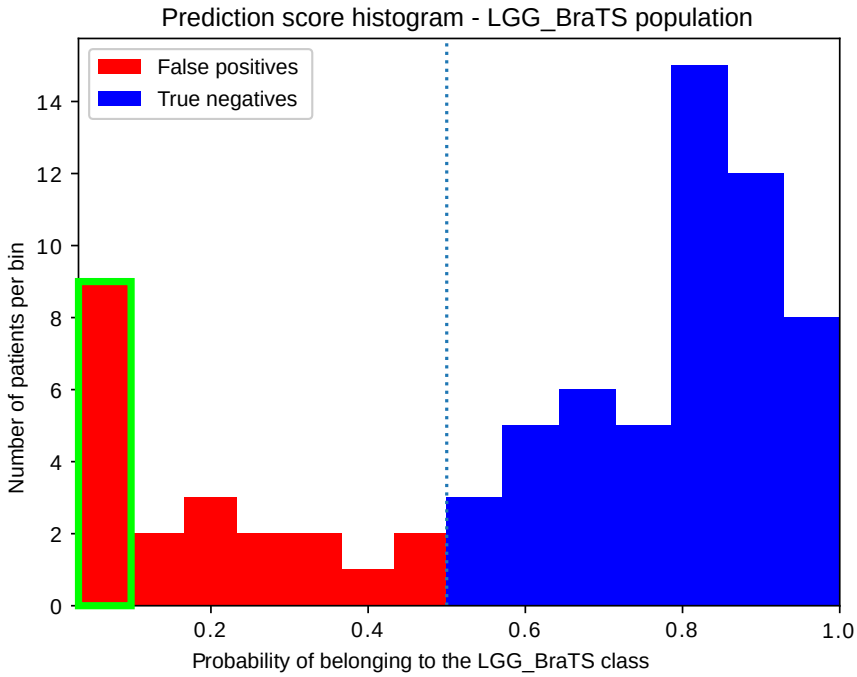

Figure 1a: Prediction score histogram of the $\text{LGG}_{\text{BraTS}}$ patients. False positives below the 0.5 probability threshold are shown in red, while true negatives above 0.5 are shown in blue. The most confident false positive patients (left) show radiological consistencies with HGG_{WHO} tumors. (see Fig. 1b)

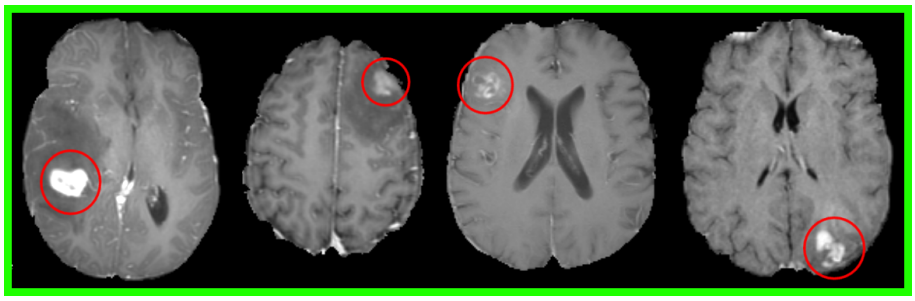

Figure 1b: 4 of the 9 most false positive patients of the prediction histogram. They all show radiomic biomarkers found in the HGG_{WHO} group, such as gadolinium enhancement and/or necrotic cavities.

3.4. Radiomics on false positives and negatives

To study the consistency of prediction, we can associate the prediction score of a patient and its MR image appearance. If a patient shows many radiological criteria of being a HGG_{WHO} and has been classified as HGG_{BraTS} by our classifier with a high probability, then our classifier worked correctly. The same reasoning can be applied to a LGG_{WHO} patient with a high LGG_{BraTS} response.

Analyzing the prediction score of each patient gives us information about how close our classifier was to get the correct answer. We can link the prediction score to the visual aspect of the images. Most of our false positives show numerous criteria that would normally make them belong to the HGG_{WHO} group. Four examples of gadolinium-enhancing false positives are shown in (Fig. 1b). These patients show gadolinium-enhancement and/or necrosis, which are radiological signs of HGG_{WHO}. This illustrates that our classifier, trained to analyze the image with the same features used by radiologists to discriminate between LGG_{WHO} and HGG_{WHO} tumors, is able to discriminate subgroups inside the LGG_{BraTS} population corresponding to LGG_{WHO} and HGG_{WHO} groups. In this case, an external observer can identify radiologically HGG_{WHO} patients among the false positives of our classifier. We can also suppose that the HGG_{BraTS} group used for training is strongly representative of "high gradeness", as it is composed of glioblastoma multiform. This caused our classifier to groups together the *highest* grade patients. Yet, our classification frontier cuts through the LGG_{BraTS} by separating radiologically-looking LGG_{WHO} and HGG_{WHO}. This may be explained by the features we used, which are based on the radiological analysis for WHO grade classification.

4. A new groundtruth for BraTS

4.1. Label gathering with expert radiologists

We want to see if our radiological criteria match the expert analysis. This is why we created a labelling task for expert radiologists. Using a Python module in Slicer3D [28][29], we were able to present each patient in randomized order.

We chose to present 100 patients so that the task would last about an hour. This allowed us to present all 75 $\text{LGG}_{\text{BraTS}}$ patients, with 25 $\text{HGG}_{\text{BraTS}}$ patients as complement. An odd number of 5 experts analyzed the images and chose to assign each patient in the LGG_{WHO} or HGG_{WHO} category. Each of them then filled a list of criteria to later compare homogeneity of the responses. Using this list shows that our radiologists used necrosis and contrast enhancement as their main criteria, then relied on various details such as intralesional bleeding, FLAIR inhomogeneity or mass effect for a finer classification. As we asked an odd number radiologists to participate in this task, we used majority voting as the decision rule to create new groundtruth data coherent with the radiological analysis.

Details about the voting distribution are shown in Table 4. We can see that the radiologists reached a 5/5 consensus in 23 cases out of 25 when labelling a $\text{HGG}_{\text{BraTS}}$ patient. Only 2 cases were with a "4 votes against 1" situation. These high numbers can be explained because the $\text{HGG}_{\text{BraTS}}$ patients are glioblastoma multiform cases, which are radiologically very different from LGG_{WHO} patients. This consistency of the $\text{HGG}_{\text{BraTS}}$ group also explains why no label was changed by the voting process. Only 49 out of 75 cases of $\text{LGG}_{\text{BraTS}}$ patients reached complete consensus from the experts, while 15 cases gained 4 votes out of 5; and 11 cases were more ambiguous, receiving only 3 votes out of 5. Our experts described these ambiguous cases as patients on the edge of anaplasic transformation and were more inclined to label them with the high grade status in order to start intensive care without delay.

Majority voting changed the grade label from $\text{LGG}_{\text{BraTS}}$ to HGG_{WHO} of 44 patients out of 75. For 29 patients, these labels were changed after a 5/5 consensus; 8 after a 4/5 vote; and 7 after a 3/5 vote. We can note than more than half of $\text{LGG}_{\text{BraTS}}$ patients changed label, which can raise questions about the quality of classification produced by previous works published with this dataset, as real LGG_{WHO} patients appear to be scarse. LGG_{WHO} patients can be asymptomatic and therefore under-diagnosed, which explains the difficulty of creating a large LGG_{WHO} dataset.

Table 4: Radiologist groundtruth voting analysis

	LGG_{BraTS}			HGG_{BraTS}			Whole test		
Majority voting	5/5	4/5	3/5	5/5	4/5	3/5	5/5	4/5	3/5
# of cases	49	15	11	23	2	0	72	17	11
Label changed	44/75			0/25			44/100		

4.2. Evaluating our classifier on our proposed groundtruth data

This groundtruth data gives us new groups of radiologically-looking LGG_{WHO} and HGG_{WHO}, but with an important class imbalance. After majority voting, our new groundtruth only has 31 LGG_{WHO} patients and 254 HGG_{WHO} patients. As a proof of concept, we wanted to see if we could still learn efficiently on this new groundtruth or if the class imbalance would impact the performances. Using the same features and parameters, we trained a new SVM classifier on this radiological groundtruth. Despite this important class imbalance, we reached similar accuracy, slightly lower sensitivity and specificity, which shows that our learning method seems robust to class imbalance. Detailled results are shown in Table 5. The confusion matrix with our radiological groundtruth is shown in Table 6. Compared to Table 3, this classifier is less accurate for LGG_{WHO} patients and more accurate for HGG_{WHO} patients.

Table 5: Performance comparison between groundtruth datasets

BraTS groundtruth			Radiologists groundtruth		
Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
84,1%	87,0%	75,9%	82,4%	83,9%	70,6%

Table 6: Confusion matrix for the SVM classifier trained on our radiological groundtruth

		Radiological groundtruth	
		LGG_{WHO}	HGG_{WHO}
SVM	LGGsvm	11	8
	HGGsvm	20	246

5. Comparing two classifiers using prediction scores

5.1. Comparing distributions with prediction score histograms

For the classifier trained on our new groundtruth, we have plotted the prediction score histogram for every $\mathrm{LGG_{WHO}}$ patient, Fig. 2. We can see on this new histogram that we only have 3 patients in the false positives close to 0. Instead, the modal bin is close to 0.3. This shift of the false positive mode to the right of the histogram reflects an improvement in the behavior of the classifier : when wrong, our classifier assign a label with a higher uncertainty. Contrary to Fig. 1a, the patients in the modal bin don't show radiological signs of being $\mathrm{HGG_{WHO}}$ patients. Therefore, despite its small number of $\mathrm{LGG_{WHO}}$ patients, we can say that our groundtruth data allow our classifier to give more radiologically consistent results.

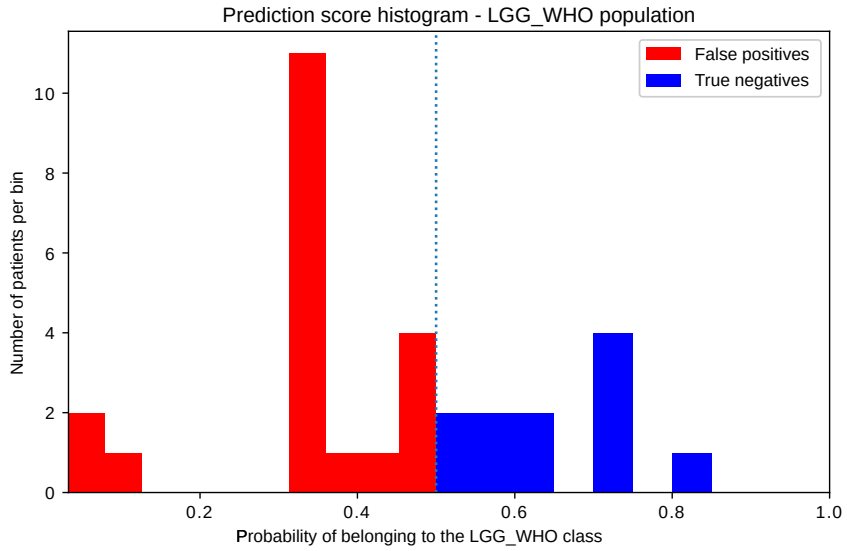

Figure 2: Prediction score histogram for the $\mathrm{LGG_{WHO}}$ patients. Compared to the training on the BraTS groundtruth data, the bin of radiologically $\mathrm{HGG_{WHO}}$ patients close to 0 has disappeared. Every patient in the modal bin near 0.3 is visually consistent with the radiological $\mathrm{LGG_{WHO}}$ class.

We can study the effect of the new groundtruth on the classification of

high grades. For comparison, we draw Fig. 3 two histograms of the prediction scores, one with the BraTS groundtruth, the other with our radiological groundtruth. We can see on both histograms that the modal bin is very close to 1.0, which shows a high level of confidence. The classifier trained on the radiological groundtruth seems more confident when classifying $\mathrm{HGG_{WHO}}$ patients, as almost all $\mathrm{HGG_{WHO}}$ patients are found in the first far-right bin. Therefore, despite showing a slightly lower accuracy in Table 5, we can see through the prediction score histogram that training on our new groundtruth gives our classifier a higher confidence. This behaviour is not visible when monitoring only the metrics from the confusion matrix. This shows us that the prediction score histogram can be used to gain a qualitative analysis of the classification. Thus, in addition to being a data analysis tool as shown in Fig. 1a, the histogram of prediction scores can also be used as a tool for comparing qualitatively the performances of two classifiers.

5.2. Comparing on the same population using crossed prediction scores

We want to see if this improvement in confidence is visible on the same population. We can enhance this comparison tool by plotting the crossed prediction scores, Fig. 4. This plot shows on each axis the prediction scores for one classifier, either trained on the BraTS groundtruth or on our proposed new groundtruth. A dotted line shows equal prediction for both classifiers. This plot can be divided in 4 quadrants, showing the different prediction outcomes for each classifier. The upper-left quadrant shows patients correctly classified by the classifier trained on our new groundtruth and wrongly classified by the classifier trained on the BraTS groundtruth. The upper-right quadrant shows when both classifiers are correct. The bottom left-quadrant shows when both classifiers are wrong. And the bottom right quadrant shows when the classifier trained on the BraTS groundtruth is correct but not the classifier trained on our new groundtruth.

On this plot, we can show the $\mathrm{HGG_{WHO}}$ population, composed of corrected $\mathrm{LGG_{BraTS}}$ and $\mathrm{HGG_{BraTS}}$ patients. By plotting individual patients with this

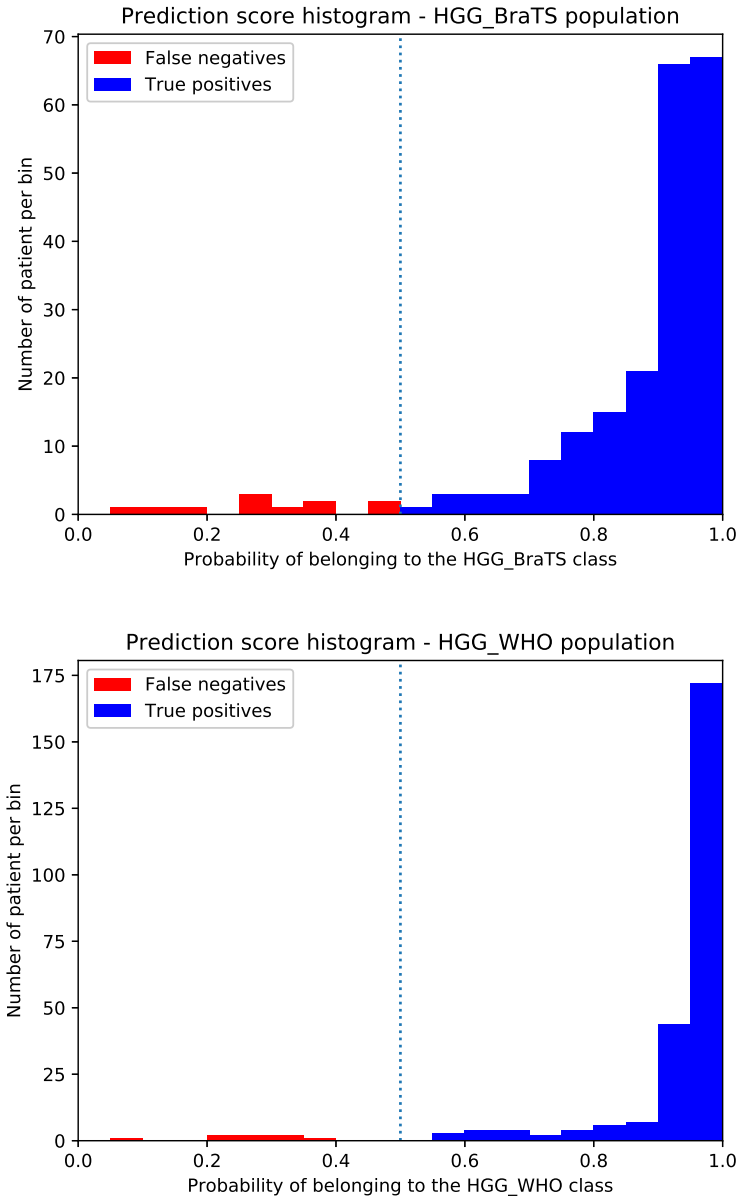

Figure 3: Prediction score histogram of $\mathrm{HGG}_{\mathrm{BraTS}}$ and $\mathrm{HGG}_{\mathrm{WHO}}$ population for a classifier trained (up) on the BraTS groundtruth and (down) on the new groundtruth. We can see that, with our new groundtruth, more patients are grouped in the far-right bin, showing improvement in the confidence of classification.

crossed comparison, we can reach an individual level of tracking. We can see that most of our corrected $\mathrm{LGG}_{\mathrm{BraTS}}$ patients are in the upper-left quadrant, sometimes reaching very high confidence in belonging to the $\mathrm{HGG}_{\mathrm{WHO}}$ class. We can also see that almost all of the $\mathrm{HGG}_{\mathrm{WHO}}$ patients are above the dotted line, showing higher confidence with our new groundtruth. This way, we can confirm that training on our new groundtruth gives better results when classifying $\mathrm{HGG}_{\mathrm{WHO}}$ patients.

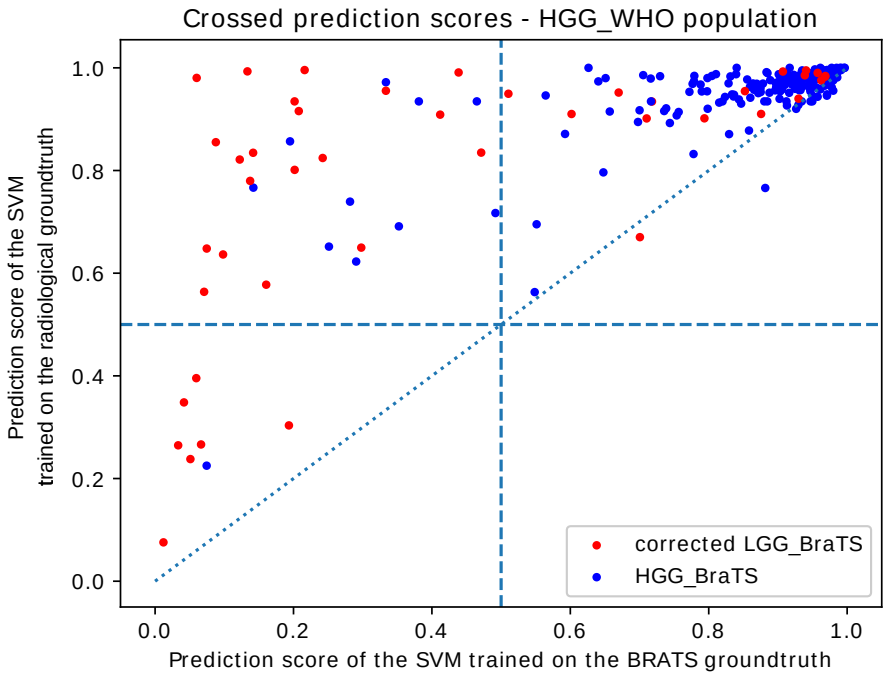

Figure 4: Crossed prediction scores of the SVM classifier trained on the BraTS groundthtruth (X-axis) and on the radiological groundtruth (Y-axis). Each point is a $\mathrm{HGG}_{\mathrm{WHO}}$ patient. The dotted line shows equal prediction scores. The dashed lines divide the space in 4 quadrants showing : (up-left) correct prediction for the classifier trained on the new groundtruth and incorrect prediction when trained on the BraTS groundtruth, (up-right) both classifiers give correct predictions, (down-left) both classifiers give incorrect prediction, (down-right) the classifier trained on BraTS gives a good answer while the classifier trained on the new groundtruth is wrong. Most of the changed $\mathrm{LGG}_{\mathrm{BraTS}}$ patients are in the upper-left corner.

6. Discussion and conclusion

In this paper, we showed how prediction scores on the BraTS dataset can be used to identify groups within the classifier performances. Studying the distribution of each patient along the prediction axis can give more information than usual evaluation based on accuracy, sensitivity and specificity. In our example, we saw a modal bin in the false positives close to 0, expressing the high confidence of our classifier to give the wrong label to these patients. This example suggests that the use of prediction score offers a new insight in understanding the behavior of classifiers. Thus, the histogram of prediction may be used in a way to "open the black box", as it is a tool of results visualization.

Detailled radiological analysis shows that our false positive patients present radiological high grade criteria such as necrosis and contrast enhancement. We asked 5 expert radiologists to label each patient allowing us to generate a new radiologically WHO-based groundtruth data for the BraTS dataset.

A new training on the radiologically coherent groundtruth shows improvement in the false positive distribution and slightly lower general performances, despite a important class imbalance. Due to the important class imbalance, our classifier trained on this groundtruth is more accurate for $\mathrm{HGG_{WHO}}$ patients and less accurate for $\mathrm{LGG_{WHO}}$ patients. Using data augmentation techniques to generate more $\mathrm{LGG_{WHO}}$ data may reduce the class imbalance and improve the $\mathrm{LGG_{WHO}}$ accuracy. Again, we used the prediction score histograms and a crossed prediction scores plot to show that, with an individual tracking level, despite these lower performances, our classifier trained on our new groundtruth was more confident in classifying $\mathrm{HGG_{WHO}}$ patients. Thus, we showed how a crossed prediction scores plot can be used as a comparison tool for classifiers analysis.

Yet, caution is to be taken for glioma grading when relying on anatomical sequences. Accessing the biological reality of the patient must involve a more complete screening with biopsy and multiparametric MRI scan. Perfusion imaging and MR Spectroscopy give the radiologist more information for grade

classification. For example, anatomical imaging has trouble identifying a non-enhancing high grade patient, so our classification can only rely on a limited body of evidence. That's why our groups based on anatomical imaging can only get close to a WHO-defined system, without the certainty and specificity of the real diagnosis.

Perspectives include creating a large multimodal dataset, with MR spectroscopy, diffusion and perfusion imaging to improve the classification. At the same time, switching to ultra high field imaging, from 3 to 7 Tesla would enhance the image quality. A WHO-based groundtruth should also allow discrimination between each 4 grades and not a binary classification. Those improvements could lead to better tools for non-invasive screening and ultimately, automatic virtual biospy.

7. Availability of groundtruth labels

The groundtruth data resulting of majority voting will be made available online. For the moment, it is available on demand to the main author.

References

[1] D. N. Louis, A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, D. W. Ellison, The 2016 world health organization classification of tumors of the central nervous system: a summary, Acta neuropathologica 131 (6) (2016) 803–820.

[2] M. B. Potts, J. S. Smith, A. M. Molinaro, M. S. Berger, Natural history and surgical management of incidentally discovered low-grade gliomas, Journal of Neurosurgery 116 (2) (2012) 365–372.

[3] M. J. van den Bent, Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective, Acta neuropathologica 120 (3) (2010) 297–304.

[4] G. Reifenberger, H.-G. Wirsching, C. B. Knobbe-Thomsen, M. Weller, Advances in the molecular genetics of gliomas—implications for classification and therapy, Nature reviews Clinical oncology 14 (7) (2017) 434–452.

[5] F. Citak-Er, Z. Firat, I. Kovanlikaya, U. Ture, E. Ozturk-Isik, Machine-learning in grading of gliomas based on multi-parametric magnetic resonance imaging at 3t, computers in biology and medicine 99 (2018) 154–160.

[6] H.-h. Cho, S.-h. Lee, J. Kim, H. Park, Classification of the glioma grading using radiomics analysis, PeerJ 6 (2018) e5982.

[7] Y. Yang, L.-F. Yan, X. Zhang, Y. Han, H.-Y. Nan, Y.-C. Hu, B. Hu, S.-L. Yan, J. Zhang, D.-L. Cheng, et al., Glioma grading on conventional mr images: a deep learning study with transfer learning, Frontiers in neuroscience 12 (2018) 804.

[8] A. K. Anaraki, M. Ayati, F. Kazemi, Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms, Biocybernetics and Biomedical Engineering 39 (1) (2019) 63–74.

[9] P. Dequidt, P. Bourdon, O. B. Ahmed, B. Tremblais, C. Guillevin, M. Naudin, C. Fernandez-Maloigne, R. Guillevin, Recent advances in glioma grade classification using machine and deep learning on mr data, in: 2019 fifth International Conference on Advances in BioMedical Engineering (ICABME), IEEE, 2019, pp. 1–4.

[10] Q. T. Ostrom, D. J. Cote, M. Ascha, C. Kruchko, J. S. Barnholtz-Sloan, Adult glioma incidence and survival by race or ethnicity in the united states from 2000 to 2014, JAMA oncology 4 (9) (2018) 1254–1262.

[11] N. Upadhyay, A. Waldman, Conventional mri evaluation of gliomas, The British journal of radiology 84 (special_issue_2) (2011) S107–S111.

[12] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal

brain tumor image segmentation benchmark (brats), IEEE transactions on medical imaging 34 (10) (2014) 1993–2024.

[13] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features, Scientific data 4 (2017) 170117.

[14] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., 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.

[15] H. Duffau, Diffuse low-grade gliomas in adults, Springer, 2017.

[16] B. L. Dean, B. P. Drayer, C. R. Bird, R. A. Flom, J. A. Hodak, S. W. Coons, R. G. Carey, Gliomas: classification with mr imaging., Radiology 174 (2) (1990) 411–415.

[17] M. Smits, M. J. van den Bent, Imaging correlates of adult glioma genotypes, Radiology 284 (2) (2017) 316–331.

[18] S. H. Patel, L. M. Poisson, D. J. Brat, Y. Zhou, L. Cooper, M. Snuderl, C. Thomas, A. M. Franceschi, B. Griffith, A. E. Flanders, et al., T2–flair mismatch, an imaging biomarker for idh and 1p/19q status in lower-grade gliomas: a tcga/tcia project, Clinical Cancer Research 23 (20) (2017) 6078–6085.

[19] A. Vamvakas, S. Williams, K. Theodorou, E. Kapsalaki, K. Fountas, C. Kappas, K. Vassiou, I. Tsougos, Imaging biomarker analysis of advanced multiparametric mri for glioma grading, Physica Medica 60 (2019) 188–198.

[20] P. Sun, D. Wang, V. C. Mok, L. Shi, Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading, IEEE Access 7 (2019) 102010–102020.

[21] V. R. Sajja, H. K. Kalluri, Brain tumor segmentation using fuzzy c-means and tumor grade classification using svm, in: smart technologies in data science and communication, Springer, 2020, pp. 197–204.

[22] K. Skogen, A. Schulz, J. B. Dormagen, B. Ganeshan, E. Helseth, A. Server, Diagnostic performance of texture analysis on mri in grading cerebral gliomas, European Journal of Radiology 85 (4) (2016) 824–829.

[23] S. Banerjee, S. Mitra, F. Masulli, S. Rovetta, Deep radiomics for brain tumor detection and classification from multi-sequence mri, arXiv preprint arXiv:1903.09240 (2019).

[24] C. Ge, Q. Qu, I. Y.-H. Gu, A. S. Jakola, 3d multi-scale convolutional networks for glioma grading using mr images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 141–145.

[25] W. Chen, B. Liu, S. Peng, J. Sun, X. Qiao, Computer-aided grading of gliomas combining automatic segmentation and radiomics, International Journal of Biomedical Imaging 2018.

[26] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer research 77 (21) (2017) e104–e107.

[27] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers 10 (3) (1999) 61–74.

[28] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, et al., 3d slicer as an image computing platform for the quantitative imaging network, Magnetic resonance imaging 30 (9) (2012) 1323–1341.

[29] 3d slicer, available at https://www.slicer.org/, last accessed sep 29 2020. URL `https://www.slicer.org/`