QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation - Analysis of Ranking Metrics and Benchmarking Results

Raghav Mehta[1], Angelos Filos[2], Ujjwal Baid[3,4,5], Chiharu Sako[3,4], Richard McKinley[6], Michael Rebsamen[6], Katrin Dätwyler[6,53], Raphael Meier[54], Piotr Radojewski[6], Gowtham Krishnan Murugesan[7], Sahil Nalawade[7], Chandan Ganesh[7], Ben Wagner[7], Fang F. Yu[7], Baowei Fei[8], Ananth J. Madhuranthakam[7,9], Joseph A. Maldjian[7,9], Laura Daza[10], Catalina Gómez[10], Pablo Arbeláez[10], Chengliang Dai[11], Shuo Wang[11], Hadrien Raynaud[11], Yuanhan Mo[11], Elsa Angelini[12], Yike Guo[11], Wenjia Bai[11,13], Subhashis Banerjee[14,15,16], Linmin Pei[17], Murat AK[17], Sarahi Rosas-González[18], Illyess Zemmoura[18,52], Clovis Tauber[18], Minh H. Vu[19], Tufve Nyholm[19], Tommy Löfstedt[20], Laura Mora Ballestar[21], Veronica Vilaplana[21], Hugh McHugh[22,23], Gonzalo Maso Talou[24], Alan Wang[22,24], Jay Patel[25,26], Ken Chang[25,26], Katharina Hoebel[25,26], Mishka Gidwani[25], Nishanth Arun[25], Sharut Gupta[25], Mehak Aggarwal[25], Praveer Singh[25], Elizabeth R. Gerstner[25], Jayashree Kalpathy-Cramer[25], Nicolas Boutry[27], Alexis Huard[27], Lasitha Vidyaratne[28], Md Monibor Rahman[28], Khan M. Iftekharuddin[28], Joseph Chazalon[29], Elodie Puybareau[29], Guillaume Tochon[29], Jun Ma[30], Mariano Cabezas[31], Xavier Llado[31], Arnau Oliver[31], Liliana Valencia[31], Sergi Valverde[31], Mehdi Amian[32], Mohammadreza Soltaninejad[33], Andriy Myronenko[34], Ali Hatamizadeh[34], Xue Feng[35], Quan Dou[35], Nicholas Tustison[36], Craig Meyer[35,36], Nisarg A. Shah[37], Sanjay Talbar[38], Marc-Andr Weber[39], Abhishek Mahajan[48], Andras Jakab[47], Roland Wiest[6,46] Hassan M. Fathallah-Shaykh[45], Arash Nazeri[40], Mikhail Milchenko1[40,44], Daniel Marcus[40,44], Aikaterini Kotrotsou[43], Rivka Colen[43], John Freymann[41,42], Justin Kirby[41,42], Christos Davatzikos[3,4], Bjoern Menze[49,50], Spyridon Bakas*[3,4,5], Yarin Gal*[2], Tal Arbel*[1,51]

[1]Centre for Intelligent Machines (CIM), McGill University, Montreal, QC, Canada, [2]Oxford Applied and Theoretical Machine Learning (OATML) Group, University of Oxford, Oxford, England, [3]Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA, [4]Department of Radiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA, [5]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, [6]Support Center for Advanced Neuroimaging (SCAN), University Institute of Diagnostic and Interventional Neuroradiology, University of Bern, Inselspital, Bern University Hospital, Bern, Switzerland, [7]Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX, USA, [8]Department of Bioengineering, University of Texas at Dallas, Texas, USA, [9]Advanced Imaging Research Center, University of Texas Southwestern Medical Center, Dallas, TX, USA, [10]Universidad de los Andes, Bogotá, Colombia, [11]Data Science Institute, Imperial College London, London, UK, [12]NIHR Imperial BRC, ITMAT Data Science Group, Imperial College London, London, UK, [13]Department of Brain Sciences, Imperial College London, London, UK, [14]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India, [15]Department of CSE, University of Calcutta, Kolkata, India, [16] Division of Visual Information and Interaction (Vi2), Department of Information Technology, Uppsala University, Uppsala, Sweden, [17]Department of Diagnostic Radiology, The University of Pittsburgh Medical Center, Pittsburgh, PA, USA, [18]UMR U1253 iBrain, Université de Tours, Inserm, Tours, France, [19]Department of Radiation Sciences, Umeå University, Umeå, Sweden, [20]Department of Computing Science, Umeå University, Umeå, Sweden, [21]Signal Theory and Communications Department, Universitat Politècnica de Catalunya, BarcelonaTech, Barcelona, Spain, [22]Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand, [23]Radiology Department, Auckland City Hospital, Auckland, New Zealand, [24]Auckland Bioengineering Institute, University of Auckland, New Zealand, [25]Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA, USA, [26]Massachusetts Institute of Technology, Cambridge, MA, USA, [27]EPITA Research and Development Laboratory (LRDE), France, [28]Vision Lab, Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529, USA, [29]EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France, [30]School of Science, Nanjing University of Science and Technology, [31]Research Institute of Computer Vision and Robotics, University of Girona, Spain, [32]Department of Electrical and Computer Engineering, University of Tehran, Iran, [33]School of Computer Science, University of Nottingham, UK, [34]NVIDIA, Santa Clara, CA, US, [35]Biomedical Engineering, University of Virginia, Charlottesville, USA, [36]Radiology and Medical Imaging, University of Virginia, Charlottesville, USA, [37]Department of Electrical Engineering, Indian Institute of Technology - Jodhpur, Jodhpur, India, [38]SGGS

Institute of Engineering and Technology, Nanded, India, [39]Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, University Medical Center, [40]Department of Radiology, Washington University, St. Louis, MO, USA, [41]Leidos Biomedical Research, Inc, Frederick National Laboratory for Cancer Research, Frederick, MD, USA, [42]Cancer Imaging Program, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [43]Department of Diagnostic Radiology, University of Texas MD Anderson Cancer Center, Houston, TX, USA, [44]Neuroimaging Informatics and Analysis Center, Washington University, St. Louis, MO, USA, [45]Department of Neurology, The University of Alabama at Birmingham, Birmingham, AL, USA, [46]Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland, [47]Center for MR-Research, University Children's Hospital Zurich, Zurich, Switzerland, [48]Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, India, [49]Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland, [50]Department of Informatics, Technical University of Munich, Munich, Germany, [51]MILA - Quebec Artificial Intelligence Institute, Montreal, QC, Canada, [52]Neurosurgery department, CHRU de Tours, Tours, France, [53] Human Performance Lab, Schulthess Clinic, Zurich, Switzerland, [54] armasuisse S+T, Thun, Switzerland.

* Senior Authors

Corresponding author: Raghav Mehta. Email - raghav@cim.mcgill.ca

Abstract

Deep learning (DL) models have provided the state-of-the-art performance in a wide variety of medical imaging benchmarking challenges, including the Brain Tumor Segmentation (BraTS) challenges. However, the task of focal pathology multi-compartment segmentation (e.g., tumor and lesion sub-regions) is particularly challenging, and potential errors hinder the translation of DL models into clinical workflows. Quantifying the reliability of DL model predictions in the form of uncertainties, could enable clinical review of the most uncertain regions, thereby building trust and paving the way towards clinical translation. Recently, a number of uncertainty estimation methods have been introduced for DL medical image segmentation tasks. Developing metrics to evaluate and compare the performance of uncertainty measures will assist the end-user in making more informed decisions. In this study, we explore and evaluate a metric developed during the BraTS 2019-2020 task on uncertainty quantification (QU-BraTS), and designed to assess and rank uncertainty estimates for brain tumor multi-compartment segmentation. This metric (1) rewards uncertainty estimates that produce high confidence in correct assertions, and those that assign low confidence levels at incorrect assertions, and (2) penalizes uncertainty measures that lead to a higher percentages of under-confident correct assertions. We further benchmark the segmentation uncertainties generated by 14 independent participating teams of QU-BraTS 2020, all of which also participated in the main BraTS segmentation task. Overall, our findings confirm the importance and complementary value that uncertainty estimates provide to segmentation algorithms, and hence highlight the need for uncertainty quantification in medical image analyses. Finally, in favor of transparency and reproducibility our evaluation code is made publicly available at https://github.com/RagMeh11/QU-BraTS.

Keywords: Uncertainty Quantification, Trustworthiness, Segmentation, Brain Tumors, Deep Learning, Neuro-Oncology, Glioma, Glioblastoma

1. Introduction

It is often challenging for machine learning groups to gain access to large-scale annotated medical imaging datasets for training and testing their algorithms. As a result, many researchers rely on smaller proprietary datasets, making it challenging to show the full potential of their algorithms and even more challenging to compare their results against other

published methods. Medical image analysis challenges (Menze et al., 2015; Bakas et al., 2018; Simpson et al., 2019; Antonelli et al., 2021; Kurc et al., 2020; Orlando et al., 2020; Codella et al., 2019; Bernard et al., 2018; Sun et al., 2021; Heller et al., 2021; Pati et al., 2021), therefore, play a pivotal role in the development of machine learning algorithms for medical image analysis by making large-scale, carefully labeled, multi-center, real-world datasets publicly available for training, testing, and comparing machine learning algorithms. In particular, the Brain Tumor Segmentation (BraTS) challenge has provided the community with a benchmarking platform to compare segmentation methods for over ten years, increasing the dataset size each year (Menze et al., 2015; Bakas et al., 2017c, 2018; Baid et al., 2021). The availability of the dataset, and the challenge itself, has permitted the development of many new successful deep learning based approaches such as the DeepMedic (Kamnitsas et al., 2016, 2017) and the nnU-Net (Isensee et al., 2018, 2021).

Despite their success in many medical image analysis challenges, the resulting deep learning algorithms are typically not translated in the clinical setting for a variety of reasons. One issue is that most deep learning models produce deterministic outputs, that is, they do not communicate the uncertainties associated with their predictions. This is problematic in the challenging context of segmenting pathological structures (e.g., tumors, lesions), as even the top performing methods produce errors. Providing uncertainties associated with the machine learning predictions could therefore permit the end-user (e.g., clinician) to review and correct the model predictions, where the model is not certain about its predictions.

Bayesian Deep Learning provides a popular framework to allow deep learning models to generate predictions and their associated uncertainties (Neal, 2012; Abdar et al., 2021). Recent advances (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Blei et al., 2017; Blundell et al., 2015) in Bayesian Deep Learning have led to widespread adaptations of the frameworks for different tasks in medical image analysis (Karimi et al., 2019; Bian et al., 2020; Dalca et al., 2018). There has been some recent work in evaluating uncertainties associated with model predictions in medical image analysis (Nair et al., 2020; Araújo et al., 2020; Ghesu et al., 2019; Tardy et al., 2019; Filos et al., 2019; Menze et al., 2020), which quantifies whether these uncertainties are indeed able to properly capture model confidence in these domains. However, to the best of our knowledge, there has not been yet a single unified approach to evaluate the model uncertainties in the context of medical image segmentation.

The main focus of this work is three-fold: i) to develop an uncertainty evaluation metric with a down-stream clinical goal in mind; ii) to benchmark the various participating teams from a recent BraTS challenge (Bakas et al., 2018), using the developed evaluation metric; and iii) to make the associated evaluation code publicly available for future benchmarking of Bayesian Deep Learning methods for medical image segmentation. In particular, we focus on developing an uncertainty evaluation criterion for brain tumor segmentation, with an end goal of developing a Computer-Aided Diagnosis (CAD) system, where the size of the pathology is small as compared to the surrounding healthy tissues. In this context, the objectives are that the uncertainty estimates associated with an automatic segmentation system reflect that the system is (a) confident when correct and (b) uncertain when incor-

rect. These criteria would mainly permit uncertain predictions to be flagged and brought to the attention of the clinical expert, rather than overburdening the expert by having to review the entirety of the prediction. To this end, here we present the resulting uncertainty evaluation metric (Mehta et al., 2020), along with the rankings and results for 14 teams participating in the Quantification of Uncertainty for Brain Tumor Segmentation (QU-BraTS) 2020 challenge. The various analyses of the methods and results produced by the different teams, highlight the necessity of the different components of our developed metric. The results indicate that the segmentation results and the associated uncertainties give complementary information as teams performing well on one task do not necessarily perform well on the other task. Qualitative results show that the developed metric measures the desired real-world properties for tumor segmentation uncertainties.

2. Related Work

In recent works (Kendall and Gal, 2017; Der Kiureghian and Ditlevsen, 2009), it has been shown that uncertainties associated with the outputs of a machine learning model are primarily divided into two sub-types: (i) *Epistemic uncertainty*, which captures the uncertainty associated with the model parameters, and (ii) *Aleatoric uncertainty*, which captures the uncertainty inherent in the data. The epistemic uncertainty captures our ignorance about which model generated our collected data. This uncertainty can be reduced to zero if the model is provided with an infinite amount of data, permitting the model parameters to learn the true distribution of the data generation model. The aleatoric uncertainty could result from measurement noise, for example, and therefore cannot be reduced even with the collection of more data. Both epistemic and aleatoric uncertainties play important roles in the field of medical image analysis. Epistemic uncertainty indicates where to trust the model output (Nair et al., 2020; Araújo et al., 2020), and aleatoric uncertainty reflects the prevalent noise in the data (Shaw et al., 2021).

Several recent papers (Ghesu et al., 2019; Nair et al., 2020; Tardy et al., 2019; Kendall et al., 2015) show cases where uncertainty estimates correlate with errors in a machine learning model. These results show promise that by estimating uncertainties, better adaptation of deep learning models in real-world scenarios is possible. However, in the medical image analysis field, to date, there is an unmet need to: (1) systemically quantify and compare how well different uncertainty estimates properly communicate the degree of confidence in the output, and (2) to rank the performance of competing estimates, given the objectives of the task and the requirements during clinical review.

The most popular metrics for measuring model confidence output are the expected calibration error (ECE) and the maximum calibration error (MCE) (Wen et al., 2020; Ashukha et al., 2020). These metrics are useful for quantitatively measuring model calibration. However, these metrics are based on the softmax probabilities. Furthermore, a simple post-processing technique like temperature scaling (Guo et al., 2017) can make a deterministic and a probabilistic model equally calibrated. ECE and MCE metrics cannot differentiate between these temperature-calibrated models.

In another paper, Lakshminarayanan et al. (2017) evaluate the usefulness of the predictive uncertainty for decision making by evaluating the model output only in cases where the model's confidence is above a user-specified threshold. Their main idea is that if the confidence estimates are well-calibrated on the data distribution seen during training, one can trust the model's predictions when the reported confidence is high and result to a different solution when the model is not confident. They showed that, indeed, when the model is evaluated on its most confident prediction, the model accuracy is high compared to when the model is evaluated on all outputs. Though this is encouraging and allows for the comparison of different uncertainty generation methods, it does not consider how many model outputs were discarded at a certain threshold. Using this criterion, a model which has low accuracy but high uncertainty when evaluated on all predictions is rewarded. However, this model is undesirable in a practical scenario, and it leads to the rejection of most of its predictions to achieve high accuracy.

Mukhoti and Gal (2018) designed a metric to quantify uncertainty for the task of semantic segmentation. They made the following assumption during the metric design: if a model is confident about its prediction, it should be accurate, which implies that if a model is inaccurate on output, it should be uncertain. With this in mind, they calculate the following two probabilities at different uncertainty thresholds: (i) p(accurate|certain): the probability that the model is accurate on its output given that it is confident; (ii) p(uncertain|inaccurate): the probability that the model is uncertain about its output given that it has made a mistake in its prediction (i.e., is inaccurate). They used the metric to compare different BDL methods for the semantic segmentation task. Though this metric is indeed useful for semantic segmentation, where each pixel in an image is labelled as one class, it is not useful for the task of pathology segmentation where there is a high class-imbalance problem and the number of pixels (voxels) of interest (pathology) is low compared to the background-healthy class. For example, in the brain tumour segmentation task, 99.9% voxels belongs to background (healthy tissue) while only 0.1% belongs to pathology. Due to high class imbalance, p(accurate|certain) would be dominated by healthy (background) voxels and majority of them can be accurately classified with high certainty.

Hall et al. (2020) developed a metric, Probability-based Detection Quality (PDQ), to evaluate the uncertainty estimate for the task of object detection. The authors combine the class labelling measure (i.e., label quality) and the bounding box detection measure (i.e., spatial quality) into the metric. Here, spatial quality measures how well the detection describes where the object is within the image. Label quality measures how effectively a detection identifies the object class. These are averaged over all possible combinations of bounding boxes and labels generated using multiple samples. The authors also organized a challenge associated with this task at the Annual Conference on Computer Vision and Pattern Recognition (CVPR) 2019. The paper and its associated challenge (Sünderhauf et al., 2019) illustrate the importance of developing uncertainty quantification metrics that are tailored to the task of interest.

Jungo et al. (2020) made the first step towards quantifying uncertainty for the brain tumor segmentation task. They compared various uncertainty generation methods such as

MC-Dropout, Deep Ensemble (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017), and others, using the standard metrics like ECE, MCE, and reliability diagrams. In addition, they proposed a new metric, Uncertainty-Error (U-E) overlap. The results showed that Deep Ensemble was able to produce more reliable uncertainty measures compared to other methods.

3. Uncertainty Evaluation Metric

This work focuses on the publicly available BraTS challenge dataset (Menze et al., 2015; Bakas et al., 2018, 2017c), which consists of both High-Grade Glioma (HGG) and Low-Grade Glioma (LGG) cases, as described in the previous BraTS manuscripts and on The Cancer Imaging Archive (TCIA) (Clark et al., 2013). However this naming convention is now obsolete and following the 2021 World Health Organization (WHO) classification of tumors of the central nervous system (CNS) (Louis et al., 2021) the data provided by the BraTS challenge should be described as including: 1) adult-type diffuse gliomas, 2) pediatric-type diffuse low-grade gliomas, and 3) pediatric-type high-grade gliomas. Note that the adult-type diffuse gliomas included in the BraTS dataset comprise both Glioblastoma (IDH-wildtype, CNS WHO grade 4) and Astrocytoma (IDH-mutant, CNS WHO grades 2-4). Ground truth labels generated and signed off by clinical experts are provided for each patient case, and consists of 3 tumor sub-regions: enhancing tumor core, necrotic core, and peritumoral edematous/infiltrated tissue (here onward referred to as edema). However, focusing on the clinical relevance, the submitted algorithms are not evaluated on each of these tumor sub-regions but on higher level tumor entities that relates directly to the surgical and radiotherapy importance. Specifically, the tumor entities considering during the evaluation and ranking of algorithms are: (i) the enhancing tumor core (ET), (ii) the tumor core (TC), which consists of the union of ET and the necrotic core, and (iii) the whole tumor (WT), which consists of all three sub-regions namely edema, necrotic core, and enhancing tumor core, and radiographically is defined by the abnormal FLAIR signal envelope. The performance of the automatic segmentation methods are finally evaluated using the Dice Similarity coefficient (referred to as DSC from here onward) and the 95th percentile of the Hausdorff distance, between the predicted labels and the provided ground truth.

The objective of the uncertainty quantification task was to evaluate and rank the uncertainty estimates for the task of brain tumor segmentation. To this end, each team provided their output labels for the multi-class segmentation task, as well as the estimated voxel-wise uncertainties for each of the associated tumor entities, namely, WT, TC, and ET. These uncertainties were required to be normalized in the range of $0-100$ for ease of computation. For each tumor entity, the uncertain voxels were filtered at N predetermined uncertainty threshold values $\tau_{1,..,N}$, and the model performance is assessed based on the metric of interest (i.e., the DSC in this case) of the remaining voxels at each of these thresholds ($\tau_{1,..,N}$). For example, $\tau = 75$ implies that all voxels with uncertainty values ≥ 75 are marked as uncertain, and the associated predictions are filtered out and not considered for the subsequent DSC calculations. In other words, the DSC values are calculated for the remaining

	DSC			
	***DSC* at 100 (baseline)**	***DSC* at 75**	***DSC* at 50**	***DSC* at 25**
Example-1	0.94	0.96	0.965	0.97
Example-2	0.92	0.955	0.97	0.975
	Ratio of Filtered TPs (1 - ($\mathbf{TP}_x$ / $\mathbf{TP}_{\text{baseline }(\tau=100)}$))			
	FTP at 100	**FTP at 75**	**FTP at 50**	**FTP at 25**
Example-1	0.00	0.00	0.05	0.1
Example-2	0.00	0.00	0.15	0.25
	Ratio of Filtered TNs (1 - ($\mathbf{TN}_x$ / $\mathbf{TN}_{\text{baseline }(\tau=100)}$))			
	FTN at 100	**FTN at 75**	**FTN at 50**	**FTN at 25**
Example-1	0.00	0.0015	0.0016	0.0019
Example-2	0.00	0.0015	0.0026	0.0096

Table 1: Change in *DSC*, Filtered True Positives (FTP) ratio, and Filtered True Negatives (FTN) ratio with change in uncertainty thresholds for two different example slices shown in Figure 1.

predictions of the unfiltered voxels. This evaluation rewards models where the confidence in the incorrect assertions (i.e., False Positives, denoted FPs, and False Negatives, denoted FNs) is low and high for correct assertions (i.e., True Positives, denoted TPs and True Negatives, denoted TNs). For these models, it is expected that as more uncertain voxels are filtered out, the *DSC* score, calculated only on the remaining unfiltered voxels, increases.

Although the criterion mentioned above helps measure performance in terms of *DSC*, the metric of interest, it does not keep track of the total number of filtered voxels at each threshold. In real practice, an additional penalty should be provided to a system that filters out many voxels at a low threshold, in order to achieve high-performance on the metric of interest, as it will increase the reviewing burden on clinical raters. One solution is to add a penalty based on the total number of filtered voxels at each uncertainty threshold. This strategy is also not ideal as it will also penalize methods that filter out FPs/FNs, areas where mistakes are made. Instead, the evaluation method chosen penalizes methods that filter out only the correctly predicted voxels (i.e., TP and TN). Given that the specific tumor segmentation task has a high-class imbalance between pathological and healthy tissues, different penalties are assigned to TPs and TNs. The ratio of filtered TPs (FTP) is estimated at different thresholds ($\tau_{1,..,N}$) and is measured relative to the unfiltered values ($\tau = 100$) such that FTP = (TP_{100} - TP_{τ}) / TP_{100}. The ratio of filtered TNs is calculated in a similar manner. This evaluation essentially penalizes approaches that filter out a large percentage of TP or TN relative to $\tau = 100$ voxels (i.e., more uncertain about correct assertions), in order to attain the reported *DSC* value, and thereby rewarding approaches with a lower percentage of uncertain TPs/TNs.

Figure 1 and Table 1 depict qualitative examples and their associated quantitative results. Here, decreasing the threshold (τ) leads to filtering out voxels with incorrect as-

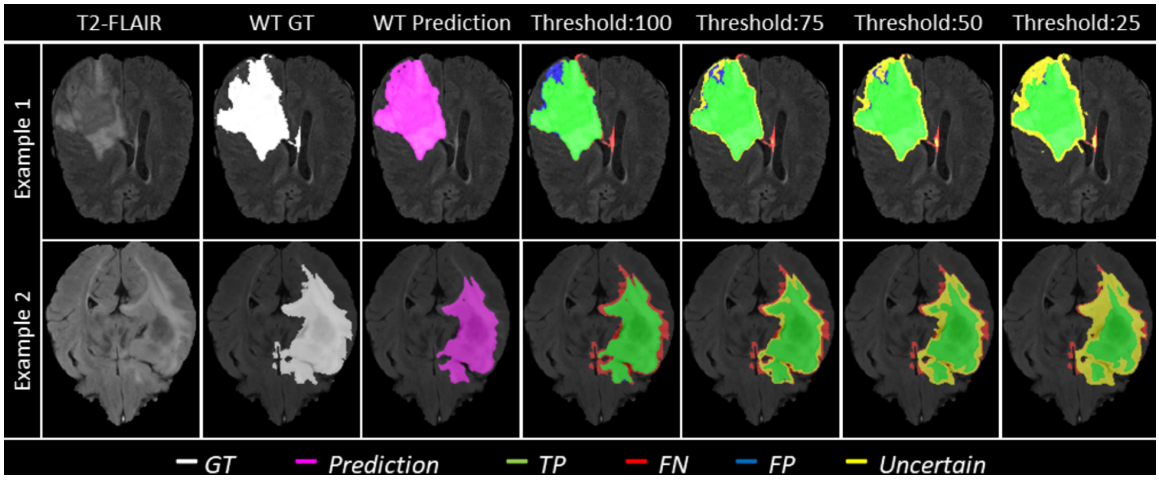

Figure 1: Effect of uncertainty thresholding on two different example patient MRI slices (Row-1 and Row-2) for whole tumor (WT) segmentation. (a) T2-FLAIR MRI. (b) WT Ground Truth (c) Overall Model Prediction (d) Results with No filtering, Uncertainty Threshold = 100. (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25. It is desired that with decrease in uncertainty threshold, more False Positives (blue) and False Negative (red) voxels are filtered out (marked as uncertain - yellow) while True Positive (green) and True Negative voxels remain unfiltered.

sertions. This filtering, in turn, leads to an increase in the DSC value for the remaining voxels. Example 2 indicates a marginally better DSC value than the slice in example 1 at uncertainty thresholds (τ) 50 and 25. However, the Ratio of FTPs and FTNs indicates that this is at the expense of marking more TPs and TNs as uncertain.

To ensure that the generated output segmentations are directly associated with the BraTS challenge protocol, the generated uncertainties are expected to be produced for these "binary" tumor entities, i.e., ET, TC, and WT. The associated uncertainties are evaluated using the metrics defined above for each tumor entity.

Finally, the resulting uncertainty measures for each team are ranked according to a unified score which combines the area under three curves: 1) DSC vs τ, 2) FTP vs τ, and 3) FTN vs τ, for different values of τ. The unified score is calculated as follows:

$$\text{score}_{\text{tumor\_entity}} = \frac{AUC_1 + (1 - AUC_2) + (1 - AUC_3)}{3}. \tag{1}$$

In the context of the BraTS uncertainty evaluation task (QU-BraTS), the score is estimated for each individual tumor entity separately and then used for ranking the participating methods.

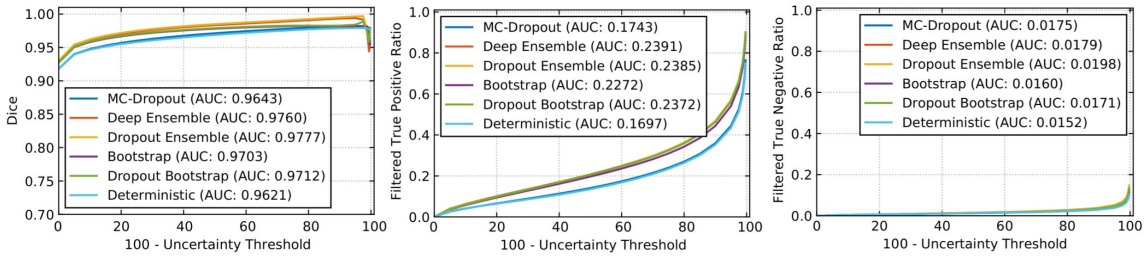

Figure 2: Effect of changing uncertainty threshold (τ) on WT for entropy measure. Specifically, we plot (left) *DSC*, (middle) Filtered True Positive Ratio, and (right) Filtered True Negative Ratio as a function of 100 - τ. We plot the curves for six different uncertainty generation methods, namely, MC-Dropout, Deep Ensemble, Dropout Ensemble, Bootstrap, Dropout Bootstrap, and Deterministic. All methods use entropy as a measure of uncertainty.

3.1 A 3D U-Net Based Experiment

A set of experiments were devised, in order to show the workings of the derived uncertainty evaluations and rankings. A modified 3D U-Net architecture (Çiçek et al., 2016; Mehta and Arbel, 2018) generates the segmentation outputs and corresponding uncertainties. The network was trained ($n = 228$), validated ($n = 57$), and tested ($n = 50$) based on the publicly available BraTS 2019 training dataset ($n = 335$) (Menze et al., 2015; Bakas et al., 2018, 2017c,a,b). The performances of WT segmentation with the entropy uncertainty measure (Gal et al., 2017), which captures the average amount of information contained in the predictive distribution, are shown in Figure 2. Here uncertainties are estimated using MC-Dropout (Gal and Ghahramani, 2016), Deep Ensemble (Lakshminarayanan et al., 2017), Dropout Ensemble (Smith and Gal, 2018), Bootstrap, Dropout Bootstrap, and a Deterministic softmax entropy measure. Dropout bootstrap shows the best *DSC* performance (highest AUC), and has the worst performance for FTP and FTN curves (highest AUC). This result shows that the higher performance in *DSC* is at the expense of a higher number of filtered correct voxels. Overall, the metric is working in line with the objectives. However, there is no clear winner amongst these uncertainty methods in terms of rankings.

4. BraTS 2020 Quantification of Uncertainty (QUBraTS) challenge – Materials and Methods

4.1 Dataset

The BraTS 2020 challenge dataset (Menze et al., 2015; Bakas et al., 2018, 2017c,a,b) is divided into three separate cohorts: Training, Validation, and Testing. The Training dataset is composed of multi-parametric MRI (mpMRI) scans from 369 diffuse glioma patients. Each mpMRI set contains 4 different sequences: native T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid-Attenuated-Inversion-Recovery (FLAIR). Each MRI volume is skull-stripped (also known as brain extraction) (Thakur et al., 2020),

co-aligned to a common anatomical atlas (i.e., SRI24 (Rohlfing et al., 2010)), and resampled to $1mm^3$ voxel resolution. Expert human annotators provided GT tumor labels, consisting of 3 classes as described previously. Note that there is no "ground-truth" uncertainty label.

The BraTS 2020 Validation cohort is composed of 125 cases of patients with diffuse gliomas. Similar to the training dataset, this also contains 4 different mpMRI sequences for each case. The validation dataset allows participants to obtain preliminary results in unseen data, in addition to their cross-validated results on the training data. The GT labels for the validation data are not provided to the participants.

The BraTS 2020 Testing cohort is then used for the final ranking of the participating team. It is comprised of a total of 166 cases. The exact type of glioma is not revealed to the participating teams. Each team gets a window of 48 hours to upload their results to the challenge's evaluation platform (`https://ipp.cbica.upenn.edu/`) (Davatzikos et al., 2018).

4.2 Evaluation framework

The University of Pennsylvania's Image Processing Portal (`https://ipp.cbica.upenn.edu/`) is used to quantitatively evaluate all BraTS participating algorithms. This portal allows both the registration of new teams to get access to the BraTS datasets, and the framework for automatically evaluating all participating algorithms on all three (i.e., training, validation, and testing) cohorts[1]. In addition to the IPP, and in favor of reproducibility and transparency, we make the implementation of the quantitative evaluation of uncertainty publicly available through GitHub[2]. As mentioned previously, the evaluation framework expects the challenge participants to provide multi-class brain tumor segmentation labels and their associated voxel-wise uncertainties for three tumor entities: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). These uncertainties are expected to be normalized between 0-100 for ease of computation.

4.3 Participating Methods

In total, 14 teams participated in the QU-BraTS 2020 challenge. All teams utilized a Convolutional Neural Network (CNN) based approach for the tumor segmentation task and the generation of associated uncertainty maps. Detailed descriptions of 12/14 proposed approaches are given below[3]. Details regarding the CNN segmentation architectures utilized by each team are not described in detail here, as the focus of this paper is on uncertainty generation methods rather than on the segmentation itself. Readers are requested to refer to the individual papers (as cited below) associated with each team for more details about the CNN architecture used for the segmentation task. Note that a preliminary version of the QU-BraTS challenge was ran in conjunction with the BraTS 2019 challenge. Appendix

1. Access to the BraTS testing datasets is not possible after the conclusion of the challenge.
2. `https://github.com/RagMeh11/QU-BraTS`
3. Two teams, namely Frankenstein (Duncan et al., 2021) and NSU-btr (Groza et al., 2021), withdrew from participating in this paper.

B provides details about the participating teams and their performance. We didn't include analysis results of the QU-BraTS 2019 challenge, as the task was ran as a preliminary task without employing any statistical significance analysis ranking scheme to evaluate the participating teams.

4.3.1 Method-1: Team SCAN (McKinley et al., 2021)

The method uses the DeepSCAN (McKinley et al., 2019) model. The training of the model was performed using a combination of focal loss (Lin et al., 2020) and a Kullback-Leibler divergence: for each voxel and each tumor entity, the model produces an output $p \in (0, 1)$ (corresponding to the output of a standard binary classification network) and an output $q \in (0, 0.5)$ which represents the probability that the classifier output differs from the ground truth on that tumor entity. The probability q is supervised by the label z, which is the indicator function for disagreement between the classifier (thresholded at the $p = 0.5$ level) and the ground truth. Given q, an annealed version of the ground truth is formed, $w = (1 - x) \cdot q + x \cdot (1 - q)$. Focal KL divergence between w and p is defined as follows:

$$\text{Focal}_{\text{KL}}(w||p) = (p - w)^2(w \cdot log(w) - w \cdot log(p)).$$

Final loss function is given by:

$$\text{Loss} = 0.1 \cdot \text{Focal}(p, x) + 0.9 \cdot \text{Focal}_{\text{KL}}(w||p) + 0.9 \cdot \text{BCE}(q, z).$$

An ensemble of the networks were utilized in the final output, where from different predictions, p and q were combined to a single probability $q \cdot I_{p \leq 0.5} + (1 - q)I_{p \geq 0.5}$. The final uncertainty output (denoted q above) was normalized into the range of 0 to 100: $100*(1-2q)$. The uncertainty in the ensemble can likewise be extracted as for any ordinary model with a sigmoid output x as: $100 \cdot (1 - 2|0.5 - x|)$

While this uncertainty measure gives a measure of uncertainty both inside and outside the provided segmentation, it was empirically observed that treating all positive predictions as certain, and only assigning uncertain values to only negative predictions, gives better performance on the challenge metrics.

4.3.2 Method-2: Team Alpaca (Murugesan et al., 2021)

An ensemble of three different 2D segmentation networks (Huang et al., 2017; Chen et al., 2019; Hu et al., 2020) were used. The softmax probabilities from each of the three networks were averaged to generate the final probability maps. These probability maps were used to generate the uncertainty maps for each tumor entity. This was computed by mapping the most confident prediction value to 0 and the least confident value to 100.

4.3.3 Method-3: Team Uniandes (Daza et al., 2021)

A novel deep learning architecture named Cerberus was proposed. The uncertainty maps were produced by taking the compliment of the final segmentation softmax probability maps, and rescaling them between 0 and 100.

4.3.4 Method-4: Team DSI_Med (Dai et al., 2021)

Five attention-gated U-Net models were trained. The uncertainty maps were normalised between 0 to 100 for the four nested tumorous entities. For each uncertainty map, the maximum softmax probability from the 5 models for each voxel in each entity was taken. The voxels were either part of the given nested entity or not, judging by the segmentation maps acquired from the ensemble of 5 models. The probabilities of those voxels that belong to the nested entity were inverted and multiplied by 100. The results were then rounded to get into the 0-100 range.

Double thresholds were further applied to refine the uncertainty maps. Low and high probability thresholds for each nested entity were empirically defined: WT(0.1, 0.3), TC(0.2, 0.3) ET(0.3, 0.5). For each voxel that belongs to a nested entity, the uncertainty was set to 0 when the probability was higher than the corresponding high threshold. For each voxel that belongs to the background, the uncertainty was set to 0 when the maximum probability was lower than the low threshold. Such method enabled the adjustment of the uncertainty of nested entities and the background independently.

4.3.5 Method-5: Team Radiomics_MIU (Banerjee et al., 2019a)

The method used an ensemble of three different CNNs (Wang et al., 2018; Banerjee et al., 2019b; Doshi et al., 2019) for segmentation. Different models were trained for three different tumor entities (i.e., WT, TC and ET segmentation). Three model ensembles were used, i.e., a total of nine models were trained for the task. Averaging various probabilities is one of the best and efficient ways to get a prediction of the ensemble model in classification. The uncertainty was estimated using the concept of entropy, to represent voxel-wise variance and diversity information. The resulting uncertainty values were scaled to lie between 0 to 100.

4.3.6 Method-6: Team Med_vision (Pei et al., 2021)

The method proposed self-ensemble-resUNet. The output softmax probabilities (y_{pred}) were inverted and normalized between 0-100 to obtain the uncertainty maps ($U_{\mathbf{pred}}$): $U_{\text{pred}} = 100 \cdot (1 - y_{\text{pred}})$

4.3.7 Method-7: Team Jaguars (Rosas-González et al., 2021)

The method used an ensemble of total 7 U-Net type models. To measure uncertainty, output probabilities of each models' were averaged for each label in each voxel to obtain a new probability for the ensemble. Since the model makes a binary classification of each voxel, the highest uncertainty corresponds with a probability of 0.5. Then the normalized entropy was used to get an uncertainty measure of the prediction for each voxel:

$$H = \sum_{c \in C} \frac{p_c \cdot \log(p_c)}{log(|C|)} \in [0, 1],$$

where p_c is the sigmoid output average probability of class c and C is the set of clases, (C = {0,1} in this case). These values were multiplied by 100 to normalize it between 0 and 100.

4.3.8 Method-8: Team UmU (Vu et al., 2021)

The method proposes a Multi-Decoder Cascaded Network to predict the probability of the three tumor entities. An uncertainty score, $u^r_{i,j,k}$, at voxel (i, j, k) was defined by:

$$u^r_{i,j,k} = \begin{cases} 200 \cdot (1 - p^r_{i,j,k}), & \text{if } p^r_{i,j,k} \geq 0.5 \\ 200 \cdot p^r_{i,j,k}, & \text{if } p^r_{i,j,k} < 0.5 \end{cases}$$

where $u^r_{i,j,k} \in [0, 100]^{|R|}$ and $p^r_{i,j,k} \in [0, 1]^{|R|}$ are the uncertainty score map and probability map, respectively. Here, $r \in R$, where R is the set of tumor entities, i.e. WT, TC, and ET.

4.3.9 Method-9: Team LMB (Ballestar and Vilaplana, 2021)

The method used a V-net (Milletari et al., 2016) architecture. A combination of test-time-dropout and test-time-augmentation was used for uncertainty estimation. In particular, the same input was passed through the network 20 times with random dropout and random data augmentation. The uncertainty map was estimated with the variance for each sub-region independently. Let $Y^i = y^i_1, y^i_2, ..., y^i_B$ be the vector that represents the i^{th} voxel's predicted labels, the voxel-wise uncertainty map, for each tumor entity (WT,TC,ET), was obtained as the variance:

$$\text{var} = \frac{1}{B} \sum_{b=1}^{B} (y^i_b - y^i_{\text{mean}})^2,$$

where y^i_{mean} represents the mean prediction across b samples.

4.3.10 Method-10: Team Matukituki (McHugh et al., 2021)

A multisequence 2D Dense-UNet segmentation model was trained. The final layer of this model is a four-channel soft-max layer representing the labels 'no tumor', 'edema', 'necrosis', and 'ET'. Uncertainty values were obtained from the final layer of the segmentation model for each label as follows: For WT, initial uncertainty values were obtained by adding the voxel-wise soft-max values of 'edema + necrosis + ET'. The initial uncertainty values for TC were the voxel-wise sum of 'necrosis + ET'. The initial uncertainty of the ET were the values of the voxel-wise soft-max channel representing ET. For all labels, the initial uncertainty values were clipped between 0 and 1. They were then modified according to the function: uncertainty = (1 – initial uncertainty) x 100. Finally, uncertainty values of 99 were changed to 100.

4.3.11 Method-11: Team QTIM (Patel et al., 2021)

The method used an ensemble of five networks for the estimation of voxel-wise segmentation uncertainty. Mirror axis-flipped inputs were passed through all models in the ensemble, resulting in 40 predictions per entity. These predictions were combined by directly averaging the model logits, denoted as l_x. A voxel with high predictive uncertainty will have

$|l_x| \approx 0$, whereas a voxel with high predictive certainty will have $|l_x| \gg 5$. To explicitly quantify uncertainty (U) in the range 0 (maximally certain) to 100 (maximally uncertain), the following formula is used:

$$U_x = \begin{cases} 200 \cdot \sigma(l_x) & \text{if } 0 \leq \sigma(l_x) < 0.5 \\ 200 \cdot (1 - \sigma(l_x)) & \text{otherwise} \end{cases}$$

where the σ function converts the ensembled logits to probabilities.

4.3.12 Method-12: Team Nico@LRDE

A cascade of two 3D U-Net type networks were employed for the task of brain tumor segmentation and its associated uncertainty estimation. The first network was trained for the brain tumor segmentation task. The second network was trained for predicting where the segmentation network was making wrong predictions. Here, the ground truth for training this network was generated as following: the ground truth was considered ones (present) at voxels where the segmentation network was wrong and it was considered as zeros (absent) at voxels where the segmentation network was correct. This way, the uncertainty networks learns to return zeros where the segmentation networks is generally true and values next to one where the segmentation networks will have issues to predict correctly the segmentation ground truth. The output of uncertainty estimation network (second network) was normalized between 0-100.

5. Analysis

In this section, we present the complete analyses and evaluation of teams that participated in the QU-BraTS 2020 challenge. Section 5.1 provides the description of the evaluation and ranking strategy followed during the QU-BraTS 2020 challenge. In Section 5.2.1 we provide the overall ranking results (accounting for all tumor entities) according to which the winning teams were announced at the challenge (Figure 3). In the same section, we also offer a comparison with their ranking on the segmentation task. Then, Section 5.2.2 provides the ranked order of the participating teams according to the individual tumor entities (Figure 4-6), followed by our ablation study (in Section 5.2.3) on the metrics incorporated in the general score (Equation 1) (Figure 7-9). Table 2 encapsulates a summary of the ranked order of the participating teams for all this analysis. Finally, Section 5.3 provides some qualitative results to highlight the effect of uncertainty thresholding filtering for all participating teams.

5.1 Ranking Scheme: BraTS 2020 challenge on uncertainty quantification (QU-BraTS)

The ranking scheme used during the challenge comprised the ranking of each team relative to its competitors for each of the testing subjects, for each evaluated tumor entity (i.e., ET, TC, WT) using the overall score (Equation 1). This ranking scheme led to each team ranked for 166 subjects for three regions, resulting in 498 individual rankings. For each team, first, the individual ranking for each patient was calculated by adding ranks across each region. This ranking is referred to as the Cumulative Ranking Score (CRS). For each team, the Normalized Ranking Score (NRS) was also calculated for each patient

Teams	Challenge Ranking (Section 5.2.1)		Variations					
			Ranking Based on Individual Tumor Entities (Section 5.2.2)			Ranking Based on Ablation Study (Section 5.2.3)		
	QU-BraTS Ranking (9)	Segmentation Ranking (18)	Whole Tumor (13)	Tumor Core (11)	Enhancing Tumor (11)	DSC AUC (10)	DSC AUC and FTP AUC (9)	DSC AUC and FTN AUC (12)
SCAN	1	4	1	1	1	6	2	4
UmU	2	7	3	2	2	4	3	3
DSI_Med	2	13	2	2	3	9	3	7
QTIM	3	7	4	2	3	3	4	2
Uniandes	4	15	5	3	4	8	5	6
nsu_btr	5	13	10	8	10	1	4	9
LMB	5	20	8	4	3	10	7	8
radiomics_miu	6	13	7	5	5	2	8	3
Nico@LRDE	6	18	6	6	6	7	9	5
Jaguars	6	13	5	6	6	2	8	3
Team_Alpaca	7	10	9	7	7	2	1	1
Matukituki	8	19	11	9	9	7	4	12
Frankenstein	9	18	13	11	8	6	6	11
med_vision	9	14	12	10	11	5	7	10

Table 2: Summary of team ranking for different analysis performed in this paper. We use the ranking scheme described in Section:5.1 to rank different teams. "QU-BraTS Ranking" column depicts the actual team ranking for all participating teams in QU-BraTS 2020 challenge (Section 5.2.1). In "Segmentation Ranking" column, we also report segmentation ranking for all teams that participated in the QU-BraTS challenge. Note that the segmentation ranking is across 78 teams that participated in the segmentation task during BraTS 2020. In three columns under "Ranking based on Individual Tumor Entities" (Section 5.2.2), we provide team ranking based only on one of the three different tumor entities. Similarly, we also report the team ranking based on the ablation study of our developed score in the last three columns of "Ranking Based on Ablation Study" (Section 5.2.3). Note that for each type of ranking, total number of provided ranks (given in the bracket) varies, as we provide same rank for teams that do not have statistical significant difference between their performance (Section 5.1).

by dividing their CRS with the total number of participating teams and the total number of regions. The NRS is in the range of 0-1 for each patient. The final ranking score (FRS) was calculated by averaging the cumulative rank across all patients for each participating team. Other challenges, such as the Ischemic Stroke Lesion Segmentation Challenge (ISLES - `http://www.isles-challenge.org/`) (Maier et al., 2017), use a similar ranking scheme.

Following the BraTS challenge, further permutation testing was done to determine the statistical significance of the relative rankings between each pair of teams. This permutation testing would reflect differences in performance that exceeded those that might be expected by chance. Specifically, for each team, given a list of observed patient-level Cumulative Ranks, i.e., the actual ranking described above, for each pair of teams, repeated randomly permutation (i.e., 100,000 times) of the Cumulative Ranks for each subject were performed. For each permutation, the difference in the FRS between this pair of teams were calculated. The proportion of times the difference in FRS calculated using randomly permuted data exceeded the observed difference in FRS (i.e., using the actual data) indicated the statistical

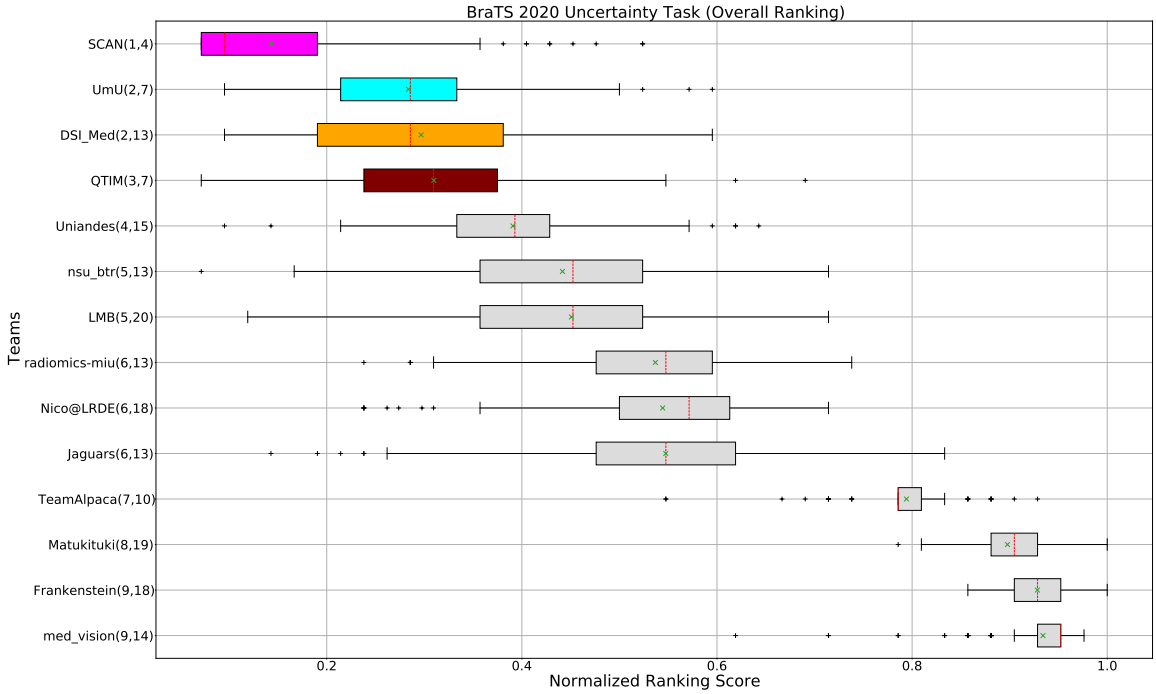

Figure 3: QU-BraTS 2020 boxplots of Normalized Ranking Score (NRS) across patients for all participants on the BraTS 2020 test set (lower is better). Boxplots for the top four performing teams are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for rest of the teams uses gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking followed by their segmentation task ranking. Note that there was no statistically significant difference between per patient ranking of teams that are ranked the same. Teams which has different ranks had statistically significant differences in their per patient ranking.

significance of their relative rankings as a p-value. Teams that do not have statistically significant difference in their FRS have similar respective rank (group) on the leaderboard[4].

5.2 Team Ranking

This section reports the final rankings of all participating teams on BraTS 2020 test dataset.

5.2.1 Overall Ranking Results

Figure 3 (and QU-BraTS ranking column in Table 2) provides a relative ranking for each team[5]. We can see that *Team SCAN* comfortably outperforms all other method and achieves the first rank in the challenge. Their Normalized Ranking Score (NRS) across all patients was ~ 0.14, while the NRS (across all patients) for the teams which achieved rank-2 (*Team UmU* and *Team DSI_Med*) was ~ 0.28. There was no statistically significant difference between *Team UmU* and *Team DSI_Med*. Thus both teams were ranked 2 in the challenge leaderboard. *Team QTIM* ranked 3rd in the challenge leaderboard and achieved marginally (though statistically significant) lower performance compared to Rank-2 teams (average NRS of ~ 0.31 compared to average NRS of ~ 0.28).

We also report the relative segmentation ranking of each team that participated in the uncertainty challenge. Note that this reported segmentation task ranking is across 78 teams that participated in the segmentation task. From the Figure 3 (and Segmentation Ranking column in Table 2), we can observe that while the *Team SCAN* (pink colour) achieves a higher ranking (Rank-4) than other teams in the segmentation task, the segmentation task ranking and the uncertainty task (QU-BraTS challenge) ranking is not the same. This is clearly visible for *Team UmU* and *Team QTIM*, as both achieved similar ranking (rank-7) in the segmentation task of BraTS 2020, while *Team UmU* was ranked second in the uncertainty task, *Team QTIM* was ranked third. Similarly, we can observe three teams that achieved Rank-13 in the segmentation task (*Team DSI_Med*, *Team nsu_btr*, and *radiomics-miu*) were ranked differently in the uncertainty evaluation task (Rank-2, Rank-5, and Rank-6, respectively). The difference in ranking across both tasks shows that performing well on the segmentation task does not guarantee good performance on the uncertainty evaluation task, and indeed both tasks are complimentary.

5.2.2 Team Ranking for individual tumor entities

As mentioned before, the BraTS challenge involves three separate tumor entities (WT, TC, and ET). The segmentation performance across these entities varies as reported in the previous BraTS challenge reports (Menze et al., 2015; Bakas et al., 2018). Specifically, the BraTS challenge reports good *DSC* across different teams for the WT segmentation task, while the performance for the ET segmentation task is relatively lower. The performance gap between different tumor entities can hinder the clinical adaptation of the segmentation algorithms. The main goal for developing the uncertainty evaluation metrics is to make algorithms more useful for clinical adaptation. Keeping this in mind, we further report the raking of each participating team according to the score (Equation 1) calculated for each tumor entity in Figure 4, Figure 5, and Figure 6.

When teams are ranked only based on their WT scores (Figure 4 and Whole Tumor column in Table 2), *Team SCAN* still comfortably outperforms other teams similar to the

4. Note that throughout the paper, we report any p-value less than 0.05 as the threshold for statistically significant differences.
5. Box plot depicting performance of each team for four different metrics - DICE_AUC, FTP_RATIO_AUC, FTN_RATIO_AUC, SCORE, for three different tumor entities - WT, TC, ET, is given in Appendix A.

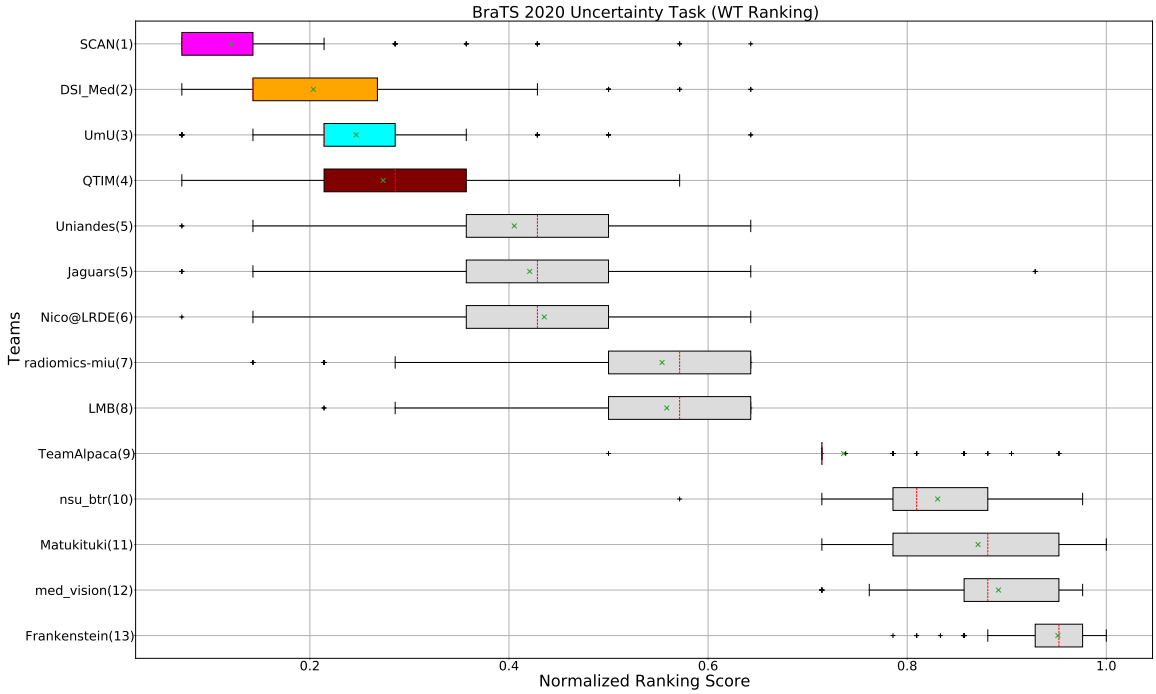

Figure 4: QU-BraTS 2020 boxplots of Normalized Ranking Score (NRS) across patients for all participants on the BraTS 2020 test set only for Whole Tumor (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3) are visualized using pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for rest of the teams uses gray colour. Here, the name of each team and their respective ranking is given on Y-Axis. Note that there was no statistically significant difference between per patient ranking of teams that are ranked the same. Teams which has different ranks had statistically significant differences in their per patient ranking.

original ranking (Figure 3). However, unlike the original ranking scheme, *Team DSI_Med* ranks statistically significantly higher compared to *Team UmU*. Similarly, from Figure 5 (and Tumor Core column in Table 2), we can observe that *Team QTIM*, *Team UmU*, and *Team DSI_Med* perform similarly to each other without any statistically significant difference between them when ranked only based on their TC score as all teams are ranked same. In Figure 6 (and Enhancing Tumor column in Table 2) *Team UmU* achieves rank-2 with statistical significance compared to *Team QTIM* and *Team DSI_Med*. We also observe no statistical significance difference between *Team QTIM*, *Team DSI_Med*, and *Team LMB*.

Overall, *Team SCAN* comfortably ranks first for all tumor entities. *Team UmU* ranks 3-2-2 for WT-TC-ET, while *Team DSI_Med* ranks 2-2-3 for WT-TC-ET. Both teams are ranked 2 when considering all tumor entities. The analysis shows that different teams achieve different ranks depending on the tumor entities, which shows that their performance differs across different tumor entities.

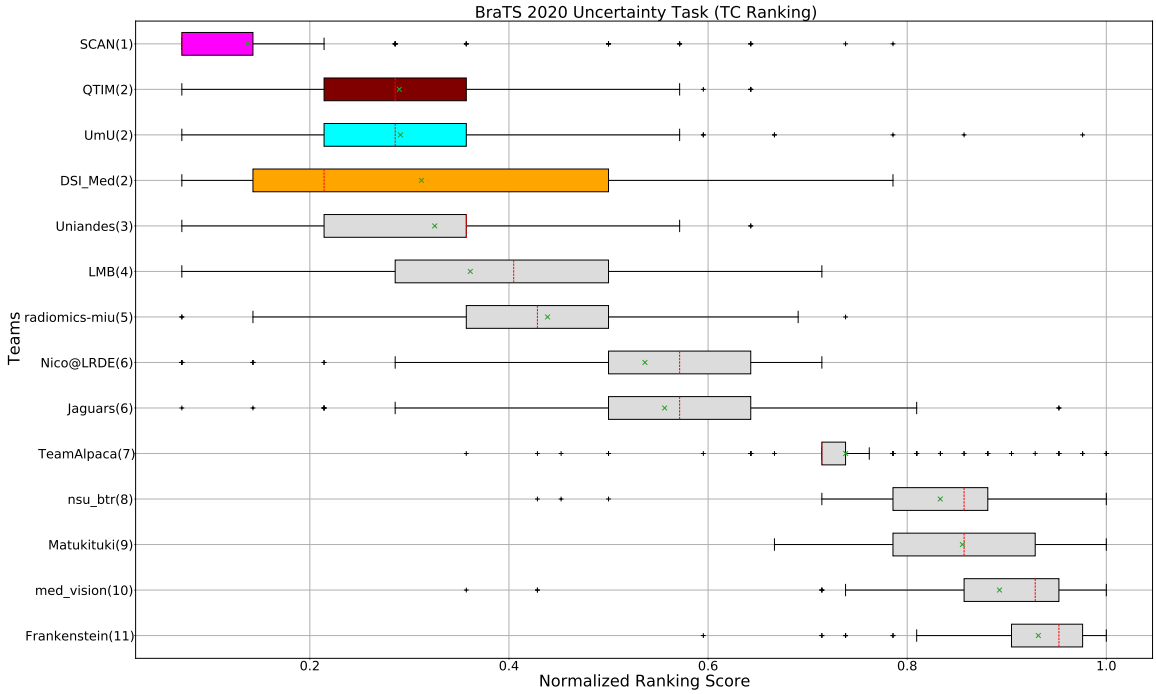

Figure 5: QU-BraTS 2020 boxplots of Normalized Ranking Score (NRS) across patients for all participants on the BraTS 2020 test set only for Tumor Core (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3) are visualized using pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for rest of the teams uses gray colour. Here, the name of each team and their respective ranking is given on Y-Axis. Note that there was no statistically significant difference between per patient ranking of teams that are ranked the same. Teams which has different ranks had statistically significant differences in their per patient ranking.

5.2.3 Ablation study on metric

The overall score for uncertainty evaluation is calculated as a combination of three different AUCs as described in Equation 1. Section 3 described the rationale behind the development of this score. As discussed previously in Section 3, evaluating the task dependent metric (in our case, *DSC*) as a function of filtered samples is critical, especially in case of pathology segmentation, where there is a high class imbalance. Recall that we expect that, by filtering more and more voxels with decrease in uncertainty thresholds, the performance on the remaining voxels measured using the task dependent metric (*DSC*) should increase but not at the expense of filtering true positive or true negative voxels. The final score consists of both the task dependent metric and filtered true positive/negatives as a function of uncertainty thresholds. In this section, we perform an ablation study of different components of the final score (*DSC*, FTP, FTN). Our analysis reaffirms that only considering one or two

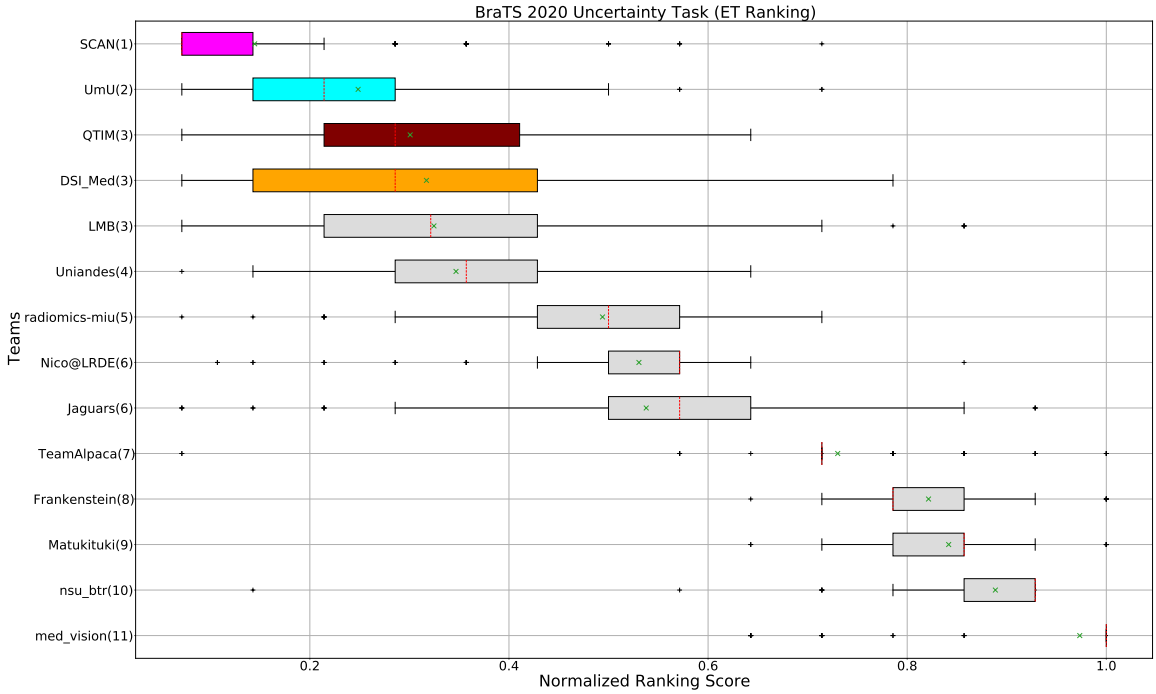

Figure 6: QU-BraTS 2020 boxplots of Normalized Ranking Score (NRS) across patients for all participants on the BraTS 2020 test set only for Enhancing Tumor (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3) are visualized using gray (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for rest of the teams uses gray colour. Here, the name of each team and their respective ranking is given on Y-Axis. Note that there was no statistically significant difference between per patient ranking of teams that are ranked the same. Teams which has different ranks had statistically significant differences in their per patient ranking.

components of the final score leads to a different ranking among participating teams.

Ranking According to *DSC* AUC: The main component of any uncertainty evaluation score is the task dependent metric, in our case, *DSC*. Many previously proposed methods for various tasks only report the value of task dependent metric at various uncertainty filtering thresholds – For example, AUC score for multiple sclerosis Nair et al. (2020). In Figure 7 (and *DSC* AUC column in Table 2), we rank participating teams according to their performance based on the AUC of *DSC* vs. Uncertainty threshold. The figure shows that higher ranking teams in this ranking scheme (*Team nsu_btr*, *Team Alpaca*, and *Team Jaguars*) are different compared to the higher ranking teams (*Team SCAN*, *Team UmU*, and *Team DSI_Med*) in the original ranking scheme (Figure 3). A closer look at the higher ranking teams according to AUC of *DSC* (Figure 7) reveals that teams like *Team Alpaca* (Section 4.3.2) achieve a good score as they use $100 - (100 \cdot \text{softmax\_confidence})$ as a proxy for uncertainty. Using an uncertainty metric which is dependent on softmax confidence

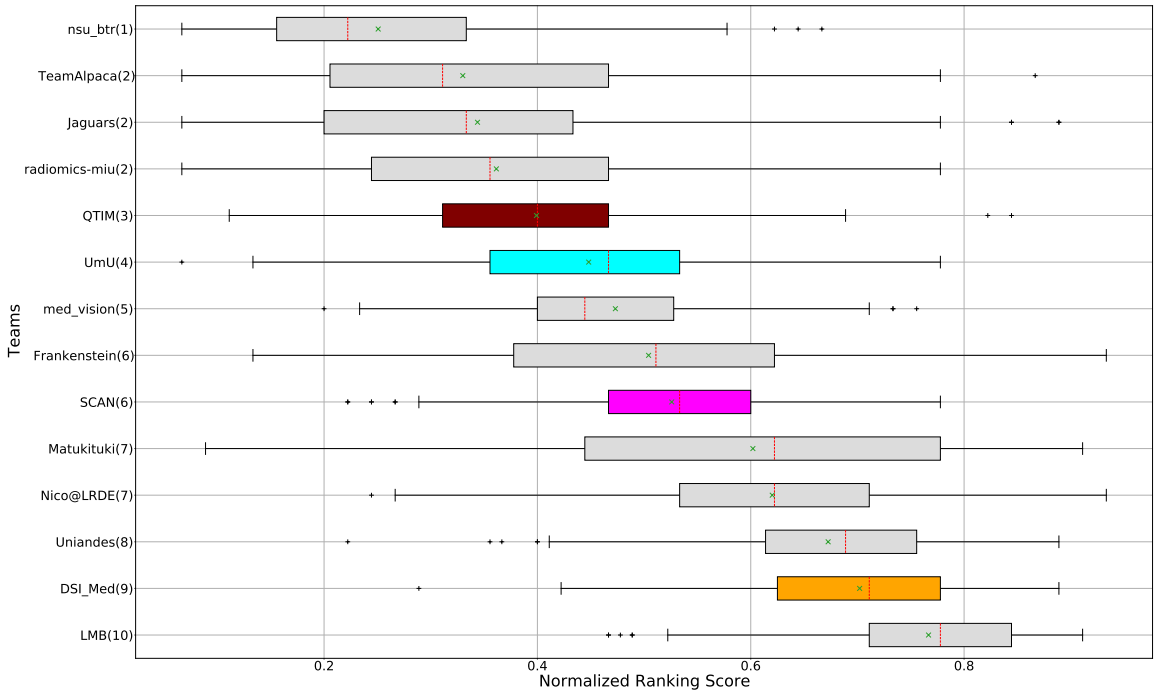

Figure 7: QU-BraTS 2020 boxplots of Normalized Ranking Score (NRS) across patients for all participants on the BraTS 2020 test set based only on DICE_AUC score (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3) are visualized using pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for rest of the teams uses gray colour. Here, the name of each team and their respective ranking is given on Y-Axis. Note that there was no statistically significant difference between per patient ranking of teams that are ranked the same. Teams which has different ranks had statistically significant differences in their per patient ranking.

leads to all background classes being marked as uncertain, increasing burden in a system where we are asking clinicians to review all uncertain voxels (Figure 14).

Ranking according to a combination of *DSC* AUC and FTP or FTN AUC: In the last section, we ranked teams according to their performance on the task dependent evaluation metrics (*DSC*) at different uncertainty thresholds. As mentioned in Section 3, the ranking team only based on their task dependent evaluation metric rewards methods which filter out many positive predictions at low uncertainty thresholds in order to attain higher performance on the metric of interest. This would increase the burden in scenarios where clinical review is needed for all uncertain predictions. To alleviate the issue, teams are ranked according to a combination of (i) AUC score for *DSC* and (ii) AUC for FTP or AUC for FTN. From Figure 8 (and *DSC* AUC and FTP AUC column in Table 2), we can conclude that a combination of both DICE_AUC and FTP_AUC alone is insufficient. It still leads to *Team Alpaca* ranked higher. As we can see in Figure 14, *Team Alpaca* marks

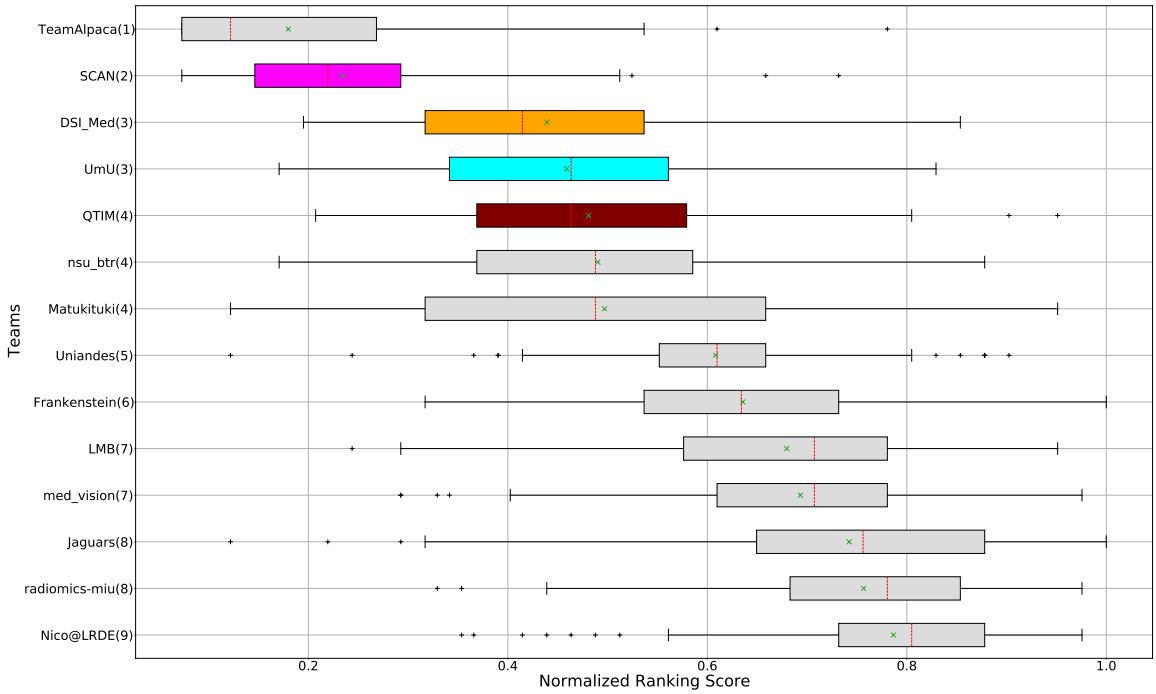

Figure 8: QU-BraTS 2020 boxplots of Normalized Ranking Score (NRS) across patients for all participants on the BraTS 2020 test set based on a combination of DICE_AUC score and FTP_AUC score (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3) are visualized using pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for rest of the teams uses gray colour. Here, the name of each team and their respective ranking is given on Y-Axis. Note that there was no statistically significant difference between per patient ranking of teams that are ranked the same. Teams which has different ranks had statistically significant differences in their per patient ranking.

all healthy-tissues (True Negative) voxels as uncertain, which reflect that the segmentation method is not confident in its prediction of healthy tissue. This is problematic as it would increase the burden in scenarios where we are expecting a clinician to review all uncertain predictions. We see a similar issue when teams are ranked only using a combination of DICE_AUC and FTN_AUC (Figure 9 and *DSC* AUC and FTN AUC column in Table 2).

Analysis in the previous two sections highlights the necessity of combining all three AUCs to calculate the final score for ranking teams in the context of uncertainty quantification of brain tumor segmentation task.

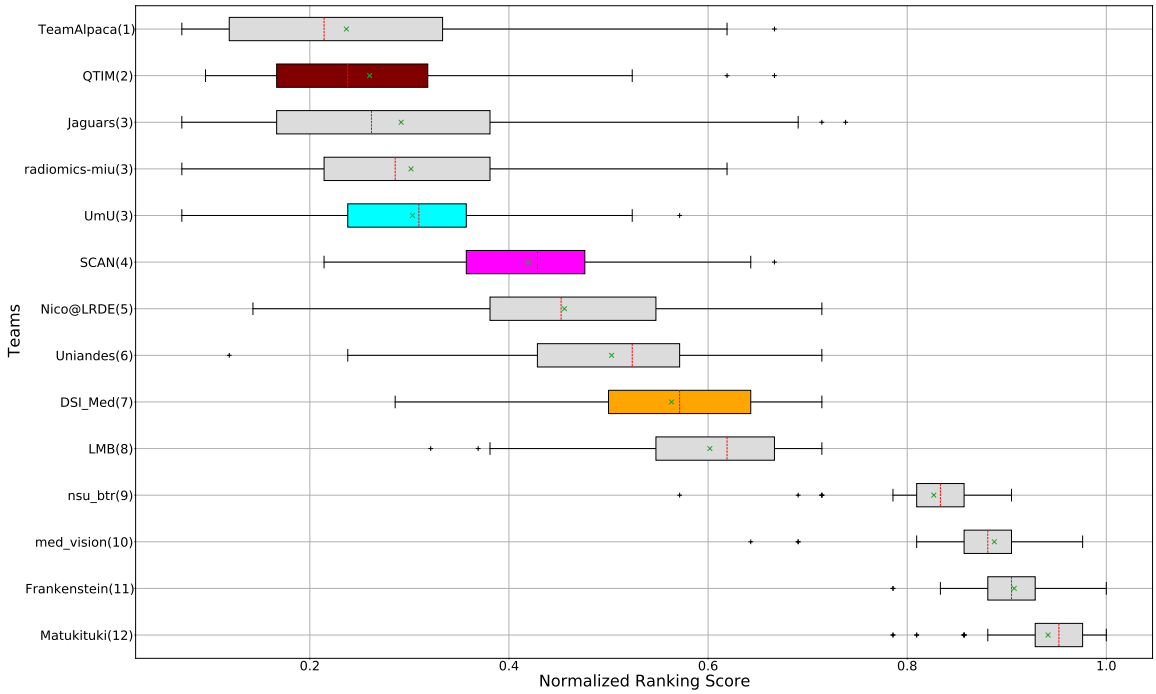

Figure 9: QU-BraTS 2020 boxplots of Normalized Ranking Score (NRS) across patients for all participants on the BraTS 2020 test set based on a combination of DICE_AUC score and FTN_AUC score (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3) are visualized using pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for rest of the teams uses gray colour. Here, the name of each team and their respective ranking is given on Y-Axis. Note that there was no statistically significant difference between per patient ranking of teams that are ranked the same. Teams which has different ranks had statistically significant differences in their per patient ranking.

5.3 Qualitative Analysis

Figure 10 - Figure 14 plots the effect of uncertainty threshold based filtering on examples slices from a few BraTS 2020 test cases for all participating teams. Green voxels represent True Positive predictions, while blue and red voxels represent False Positive and False Negative predictions, respectively. We filter out voxels at different thresholds (100, 75, 50, and 25). Filtered voxels are marked as yellow. According to the developed uncertainty evaluation metric (Section 3), we want methods that filter out (marked as yellow) false positive and false negative voxels while retaining true positive and true negative voxels as we decrease the uncertainty threshold.

In Figure 10, we visualize the effect of uncertainty based thresholding for WT segmentation on a single slice of a BraTS 2020 test case. A closer look at some of the better per-

forming teams like *Team SCAN*, *Team UmU*, and *Team DSI_Med* reveals that these teams filter out more False Positives and False Negatives at a higher threshold compared to other teams like *Team QTIM* and *Team Uniandes*. We can also observe that lower-performing teams like *Team Alpaca*, *Team Matukituki*, *Team Frankenstein*, and *Team med_vision* mark all background voxels as uncertain at a low threshold. As mentioned before, marking background voxels as uncertain is problematic as it shows that the method is not confident in its healthy-tissue segmentation and requires clinicians to review the segmentation.

In Figure 11, we plot the effect of uncertainty based thresholding for WT segmentation on another slice of the same BraTS 2020 test case. Here we observe the similar trend where higher ranked teams can filter out both False Positives and False Negatives at higher threshold than other teams. *Team SCAN* only filters negative predictions. This results in them never filtering out their False Positive predictions of the whole tumor inside the ventricles. It is problematic in a real world scenario as we do not want a method which is over-confident about its positive pathology segmentation predictions.

Figure 12 shows an example slice of a different BraTS 2020 patient and visualize the effect of uncertainty thresholding for core tumor segmentation. The figure highlights that team ranking is different across different cases as we can see that *Team SCAN* and *Team UmU* has similar prediction at *Threshold:100*. However, *Team SCAN* starts filtering out more true negatives sooner compared to *Team UmU*, which would result in *Team SCAN* ranked lower compared to *Team UmU* for this particular BraTS test case. We can observe a similar trend when comparing *Team DSI_Med* and *Team LMB*, where *Team LMB* starts filtering out more false positives sooner than *Team DSI_Med*. Similarly, in Figure 13 we can observe that in scenarios where all teams are making errors by predicting a high amount of false positives, the overall uncertainty score would be more reliant on which teams can filter out these false positives sooner. For example, *Team UmU* performs better compared to *Team DSI_Med*.

Figure 14 depicts an examples slice of uncertainty threshold based filtering for ET segmentation. Here we can see that when all teams make almost similar predictions with a high amount of true positives compared to false positives/false negatives, the overall uncertainty score is also similar across teams. Except teams which mark all background (healthy-tissue) voxels as uncertain; they perform poorly on the final score.

6. Discussion

This paper introduced a new metric (score) for evaluating uncertainties in the task of brain tumor segmentation during the BraTS 2020 challenge. The proposed score was used to rank different participating teams from the Uncertainty Quantification task of the BraTS 2020 challenge (QU-BraTS 2020).

The proposed evaluation metric was developed with the clinical objective of enabling the clinician to review only the uncertain areas of a automatic segmentation algorithm instead of the complete segmentation. Toward this end, this metric would reward algorithms that

are confidence when correct and uncertainty when incorrect. The objective was evaluated by filtering (marking as uncertain) voxels with uncertainty higher than a specified threshold as uncertain. The task-dependent DSC is measured only on the remaining unfiltered voxels. To ensure that method does not filter out a high number of correctly predicted voxels in order to achieve a better DSC, the developed evaluation score also keeps track of the number of filtered True Positive and True Negative voxels. By keeping track of these filtered TP and TN voxels, it ensures that the burden on the reviewing clinicians is not increased substantially. In short, the proposed score calculates the task-dependent metric score (i.e. DSC for segmentation), the percentage of filtered true positives and true negatives at different uncertainty thresholds, and combines them to generate a single evaluation score for a single subject.

The analysis (Section 5.2) of algorithms developed by participating team from QU-BraTS 2020 task, highlighted that the relative ranking of the participating teams for both the segmentation and uncertainty quantification tasks are different. The different ranking orders show that performing better on the segmentation task does not guarantee good performance on the uncertainty quantification task. An automatic segmentation method that provides both the segmentation and its uncertainties is more clinically relevant. Both the segmentation and the associated uncertainties provide complementary information. For example, the automatic segmentation can provide accurate results with minimal input from clinicians. In contrast, the associated uncertainty would allow clinicians to see where to trust the segmentation and where to review the segmentation before deploying it in a clinical practice.

Results in the Section 5.2.2 indicate that it is necessary to rank teams individually for each tumor entity as they rank differently across these entities. An ablation study on the proposed score (Section 5.2.3) showed the necessity of utilizing all three components (DSC, percentage of Filtered True Positive, and percentage of Filtered True Negative) for the proposed uncertainty evaluation score.

One of the significant limitations of the current performed analysis is the dependency between the segmentation results and the uncertainty generation methods, which does not allow for more in-depth analysis. It would be interesting to analyze and compare different uncertainty generation methods (e.g., *Team SCAN*, *Team UmU*, *Team Alpaca*) when the segmentation method is same across them.

We also observe a limitation of the proposed evaluation score. *Team SCAN* performs better on the overall score by not marking any positive prediction as uncertain. In a real-world scenario, a method that is always confident about its positive predictions leads to confident over-segmentation. This shows that the developed uncertainty evaluation metric is not perfect and we need to keep improving it. We observed similar trend in a recently conducted Probabilistic Object Detection challenge (Skinner et al., 2019), were the winning team attained the best score despite not using a probabilistic method. These two examples show the need to keep improving the developed task-depended uncertainty evaluation met-

ric for different tasks.

The *DSC* is a good metric for segmentation when the structure of interest contains a high number of voxels, but it is not a stable metric when calculated on a low number of voxels (Reinke et al., 2021). In the developed evaluation score, instability of the *DSC* leads to low performance at a lower threshold (more filtered voxels), as *DSC* calculation considers only a few remaining unfiltered voxels (Figure 2). The poor stability of *DSC* is a well-known issue in the literature (Reinke et al., 2021). Future work could explore a task-dependent metric that is more stable across different uncertainty thresholds (i.e., different volumes for the structure of interest).

The proposed evaluation framework evaluates uncertainties for each tumor entity as a single class segmentation/uncertainty problem, while the overall tumor segmentation is a multi-class problem. Future extensions could involve developing methods to evaluate uncertainties in the context of multi-class segmentation. Multi-class segmentation uncertainties and single-class segmentation uncertainties are different and can lead to different outcomes (Camarasa et al., 2021). In addition, the focus of the current evaluation framework is on filtering individual voxels, as most of the developed uncertainty frameworks generate per-voxel uncertainties that are not spatially correlated (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017). The recent development of spatially correlated uncertainty generation methods (Monteiro et al., 2020) points to the necessity of developing uncertainty evaluation metrics that consider the spatial correlation between pixels/voxels.

One more future direction is obtaining "ground-truth" uncertainty maps and evaluating automatic uncertainty generation methods against these maps. One recent promising direction uses inter-observer and intra-rater variation as a proxy for "ground-truth" uncertainty (Kohl et al., 2018; Baumgartner et al., 2019; Menze et al., 2020). One limitation of this approach is the assumption of attaining "ground-truth" uncertainty by asking multiple annotators to label the same images multiple times. There are many situations where a unique label cannot necessarily be attained in some regions of an image, for example at boundaries between tumor and healthy tissue in MRI due, in part, to partial volume effects. One alternative approach could be asking a single annotator to specifically mark areas they are not certain about, such as tumor boundaries in an MRI scan. These "uncertain" areas can then serve as "ground-truth", and uncertainty estimates generated by algorithms can be compared to it. Currently acquiring a "ground-truth" uncertainty is still an open area of research.

The approach developed for QU-BraTS has shown promising results in different applications. For example, the proposed score was used to evaluate uncertainties for brain tumor segmentation when a crucial MR sequence is missing (Vadacchino et al., 2021). As well, the proposed score has been used to evaluate multi-class segmentation of the carotid artery lumen and the vessel wall (Camarasa et al., 2021).

Acknowledgments

Research reported in this publication was partly supported by the Informatics Technology for Cancer Research (ITCR) program of the National Cancer Institute (NCI) of the National Institutes of Health (NIH), under award numbers NIH/NCI/ITCR:U01CA242871 and NIH/NCI/ITCR:U24CA189523. It was also partly supported by the National Institute of Neurological Disorders and Stroke (NINDS) of the NIH, under award number NIH/NINDS:R01NS042645, The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

The conflicts of interest have not been entered yet.

Appendix A - Box plots for individual scores

This appendix provides box plots for each team's performance for four different metrics (DICE_AUC, FTP_RATIO_AUC, FTN_RATIO_AUC, and Score - Equation 1) for three different tumor entities (WT, TC, and ET). The teams are ranked from better to worse performance according to mean values across all patients for each metric. Higher is better for DICE_AUC (Figure 15 - Figure 17) and Score (Figure 24 - Figure 26), while lower is better for FTP_RATIO_AUC (Figure 18 - Figure 20) and FTN_RATIO_AUC (Figure 21 - Figure 23).

Note that these box plots are different from ranking plots, as the ranking plots describe the overall performance across different tumor entities and different subjects as described in Section 5.1. From these plots, we can see that while for all three tumor entity DICE_AUC plots, *Team nsu_btr* performs better than other teams, their overall Score is comparatively lower than other teams as they do not perform well for FTP_RATIO_AUC and FTN_RATIO_AUC.

Similarly, we also observe that *Team SCAN* does not outperform other teams for DICE_AUC but comfortably outperforms other teams in FTP_RATIO_AUC. They perform relatively similar to other top-ranked teams in the FTN_RATIO_AUC metric. Overall, they achieve the best performance for the Score metric across all three tumor entities. The main reason for them outperforming other teams for FTP_RATIO_AUC is how they developed their uncertainty generation method. They found that they achieved the best results on the given Score (Equation 1) by considering all positive predictions as certain (Section 4.3.1).

In terms of overall Scores, we observe that *Team SCAN* comfortably outperforms all other teams for each tumor entity. *Team QTIM* and *Team Uniandes* report better mean scores across different patients compared to *Team SCAN*. Despite this, they do not achieve an overall better ranking for each patient, which shows the usefulness of reporting ranking and statistical-significance analysis across different patients rather than just reporting mean overall Score across patients.

Appendix B - QU-BraTS 2019

In this appendix, we analyze and briefly describe methods employed by participating teams in BraTS 2019 sub-challenge on uncertainty quantification. A total of 15 teams participated in the challenge. From these 15 teams, five teams further participated during the following year's QU-BraTS 2020 challenge.

BraTS 2019 dataset: As described in Section 3.1, BraTS 2019 dataset contains 335 patient MRIs in the training set, 125 in the validation set, and 166 in the testing set. All teams developed their method using the training set and the validation set. Ground truth segmentation for the validation set was not publicly available for the teams. The final performance of all teams was measured on the testing set, where each team had access to a 48-hour window to upload their result to the server (`https://ipp.cbica.upenn.edu/`).

QU-BraTS 2019 results on the test set: Note that we ran the task of uncertainty quantification preliminary during the challenge and did not employ any ranking scheme. Also, the metric used during the challenge was different from the one described in Section 3. Precisely, we did not calculate the AUC of Ratio of Filtered True Negatives vs. Uncertainty threshold until the validation phase was ended; and only used AUC's of *DSC* vs. Uncertainty Threshold and Ratio of Filtered True Positives vs. Uncertainty Threshold. After the validation phase, using qualitative inspection, we found that many teams were employing 1 - softmax confidence as an uncertainty measure, which is not helpful from a real clinical point of view as described in Section 3 and Section 5.3. Keeping this in mind, we added the AUC of Ratio of Filtered True Negatives vs. Uncertainty threshold during the final testing phase. Table 3 lists all team names and their performance on the BraTS 2019 test phase. The table shows that teams that employed 1 - softmax_confidence as uncertainty measure performed poorly on FTN_RATIO_AUC score (Ex. *Team Alpaca*, *Team DRAG*, *Team ODU_vision_lab*, etc.). We want to point out that we did not employ the ranking strategy used in the QU-BraTS 2020 challenge during the QU-BraTS 2019 challenge. As we discussed in Appendix A, the ranking strategy and statistical significance analysis reflect the true potential of the method compared to just ranking teams according to their mean performance across testing cases.

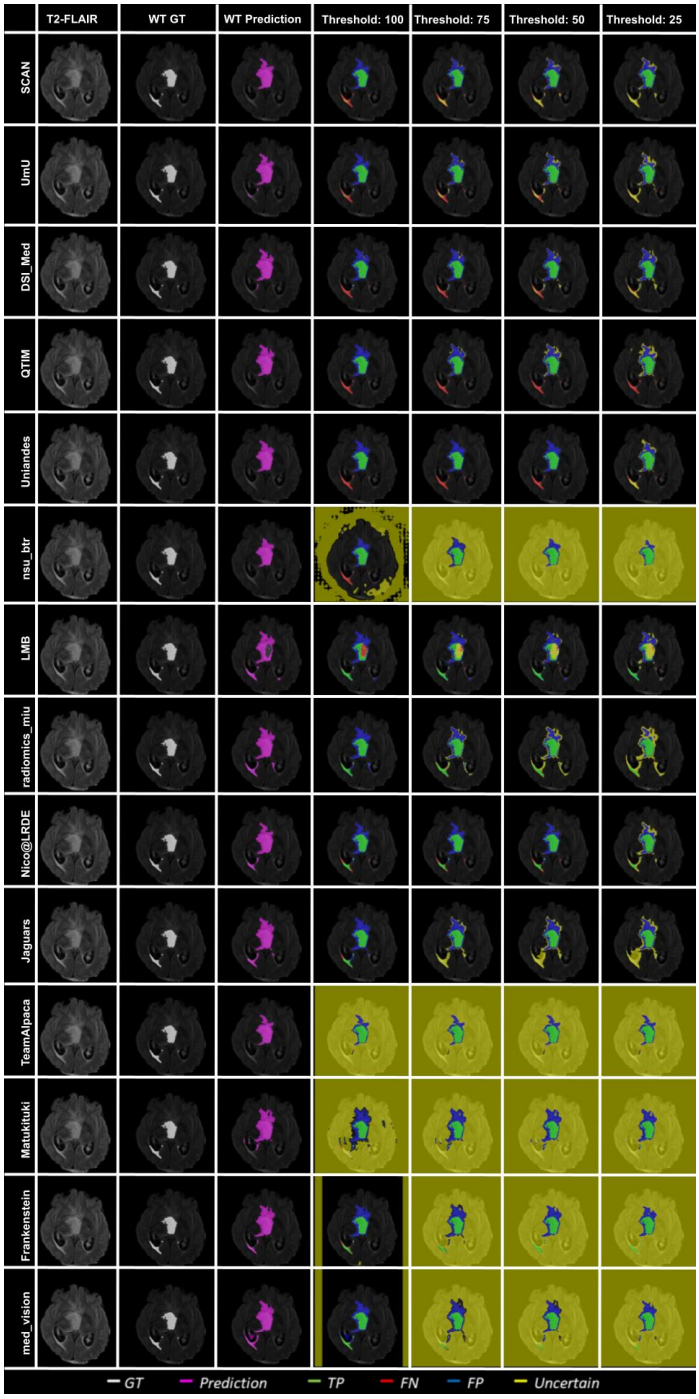

Figure 10: Effect of uncertainty thresholding on a BraTS 2020 test case for whole tumor segmentation across different participating teams. (a) T2-FLAIR MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25.

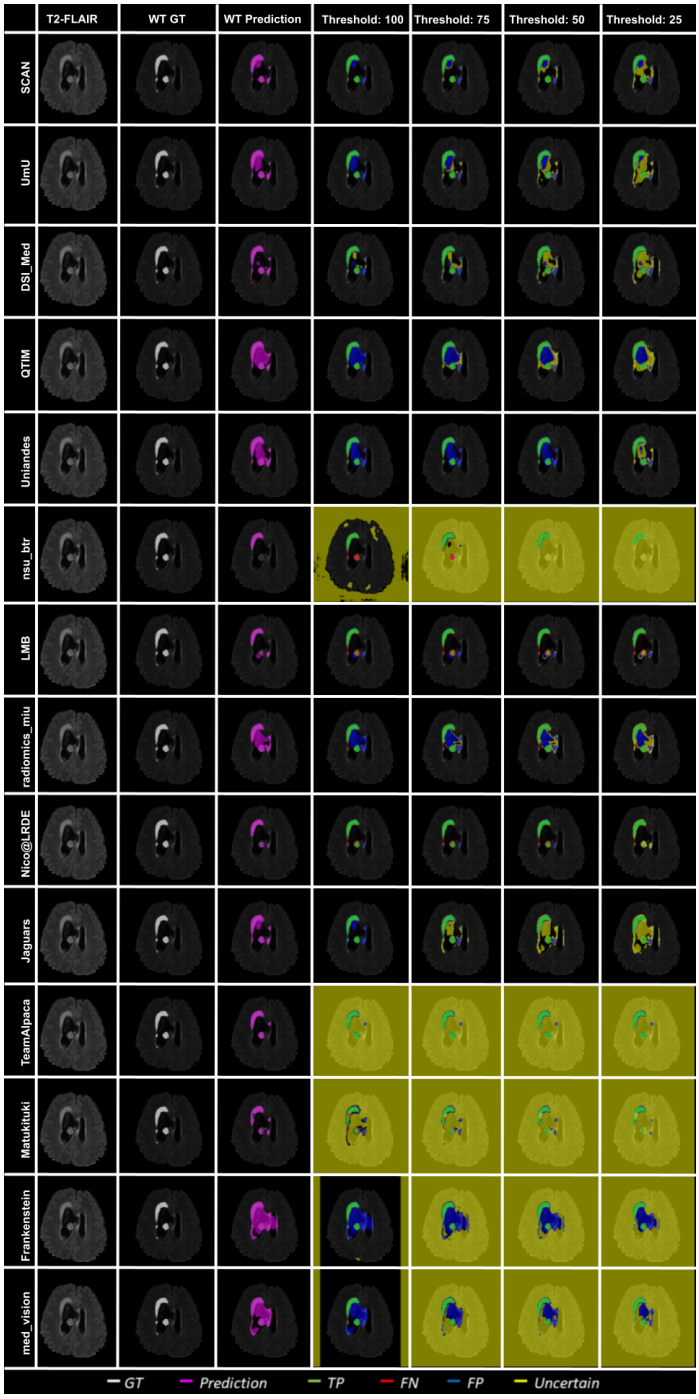

Figure 11: Effect of uncertainty thresholding on a BraTS 2020 test case for whole tumor segmentation across different participating teams. (a) T2-FLAIR MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25.

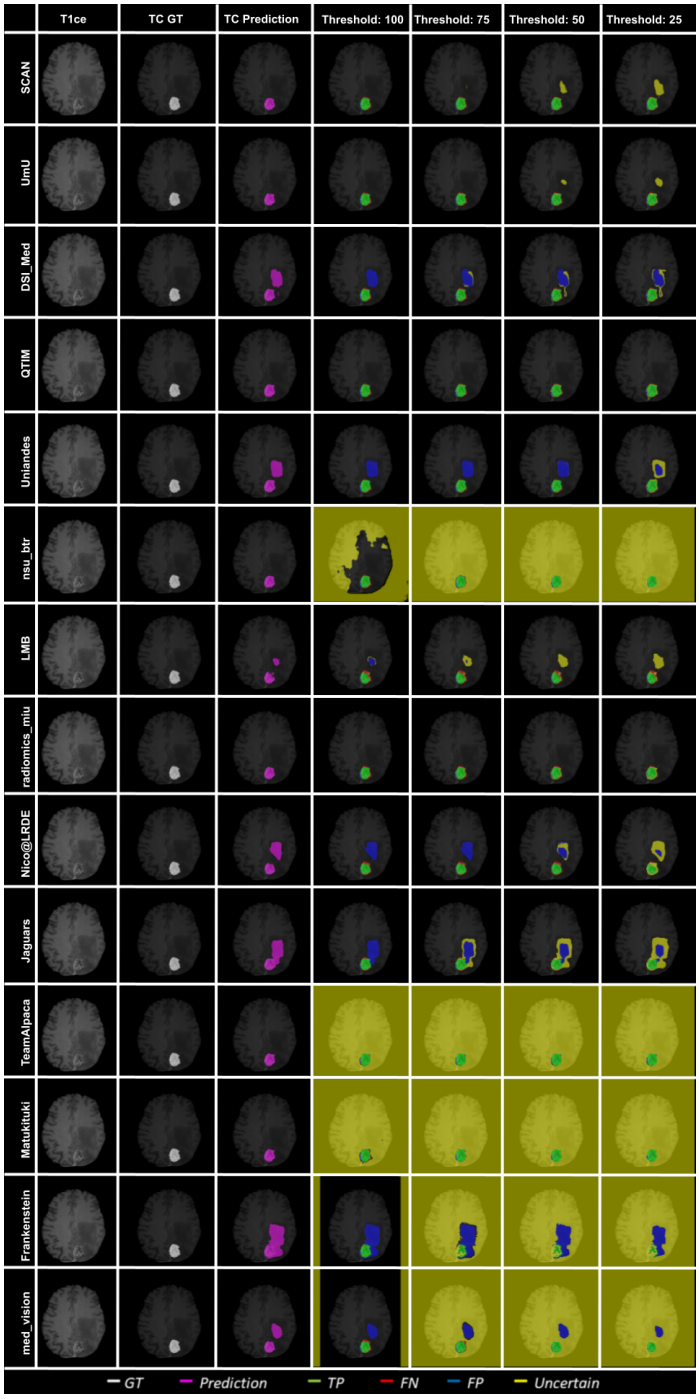

Figure 12: Effect of uncertainty thresholding on a BraTS 2020 test case for core tumor segmentation across different participating teams. (a) T1ce MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25.

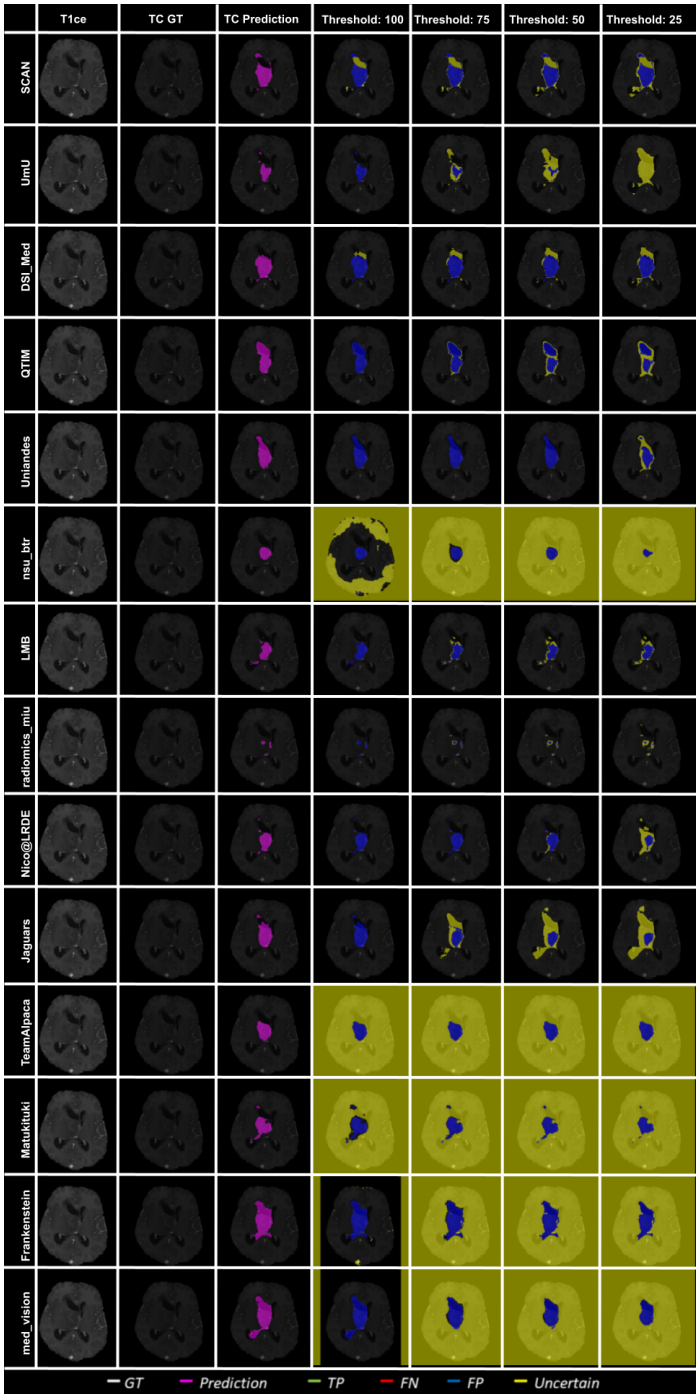

Figure 13: Effect of uncertainty thresholding on a BraTS 2020 test case for core tumor segmentation across different participating teams. (a) T1ce MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25.

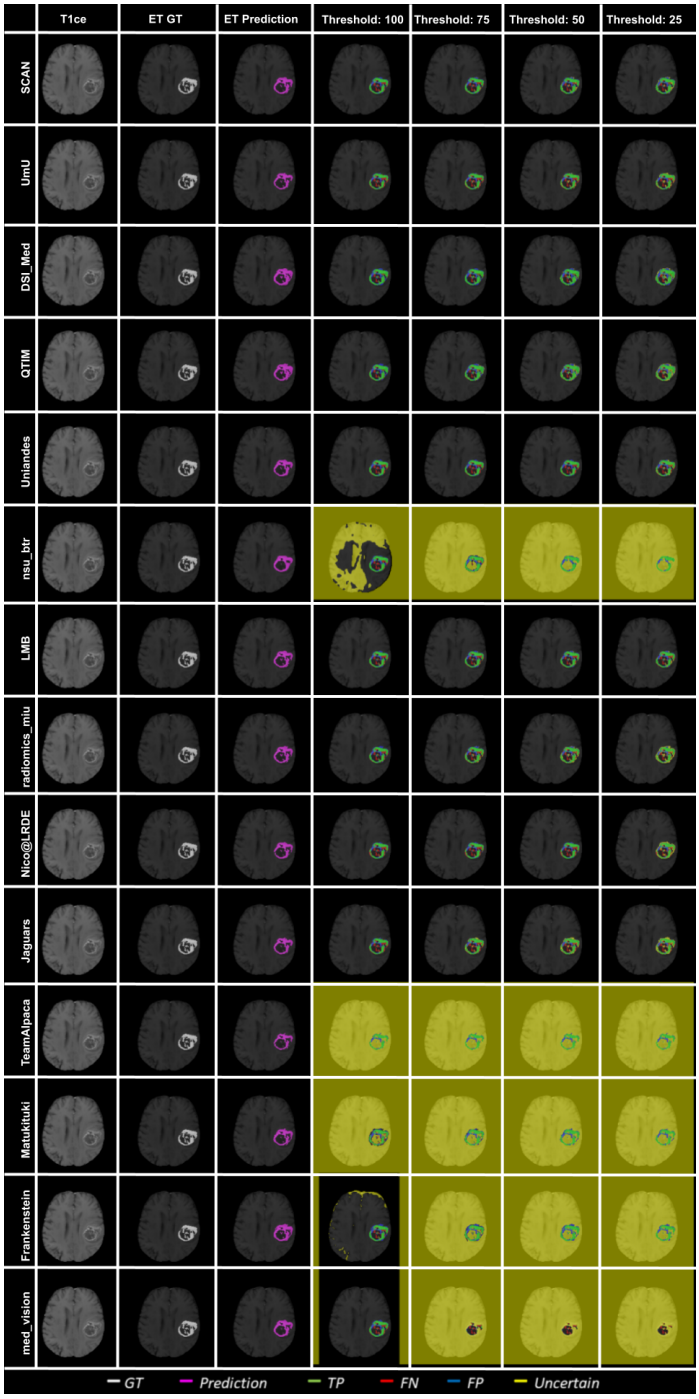

Figure 14: Effect of uncertainty thresholding on a BraTS 2020 test case for enhance tumor segmentation across different participating teams. (a) T1ce MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25.

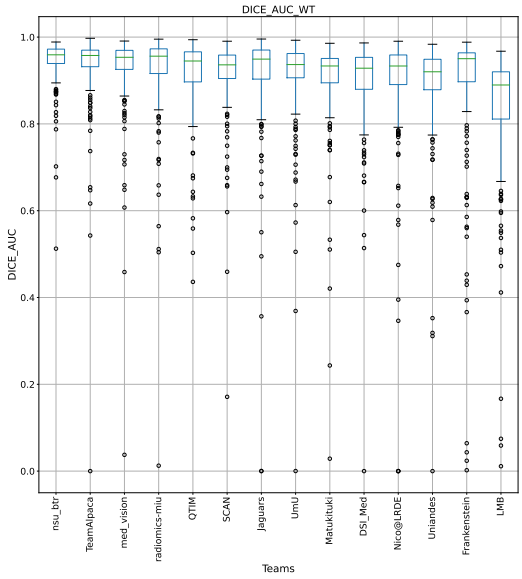

Figure 15: QU-BraTS 2020 boxplots depicting DICE_AUC distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (higher is better).

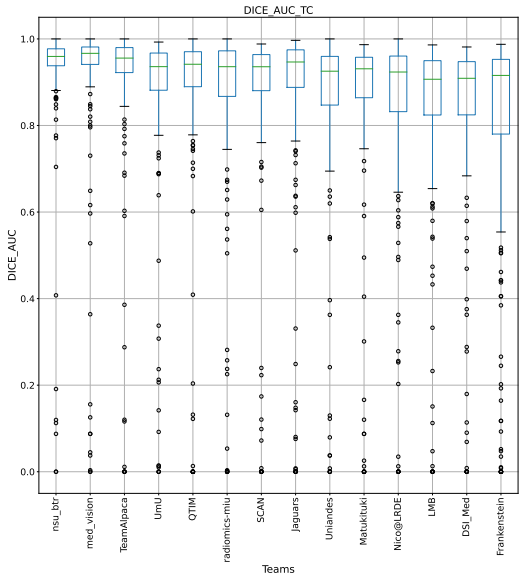

Figure 16: QU-BraTS 2020 boxplots depicting DICE_AUC distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (higher is better).

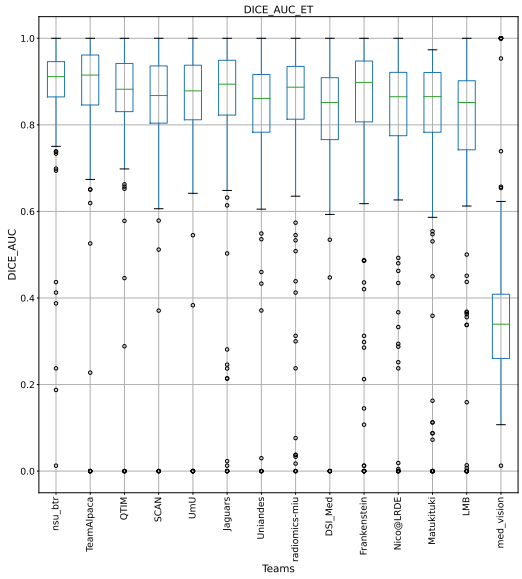

Figure 17: QU-BraTS 2020 boxplots depicting DICE_AUC distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (higher is better).

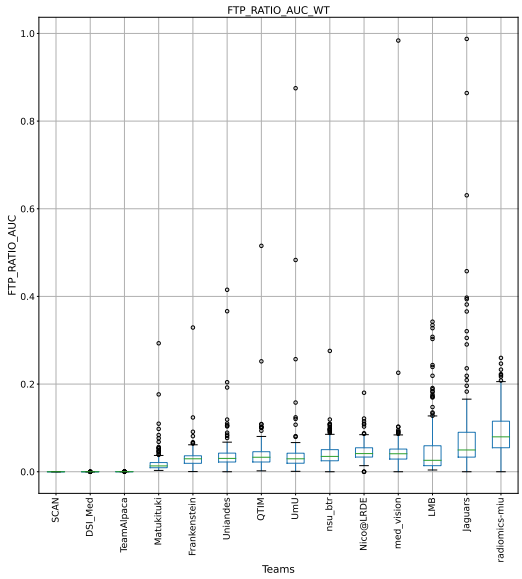

Figure 18: QU-BraTS 2020 boxplots depicting FTP_RATIO_AUC distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (lower is better).

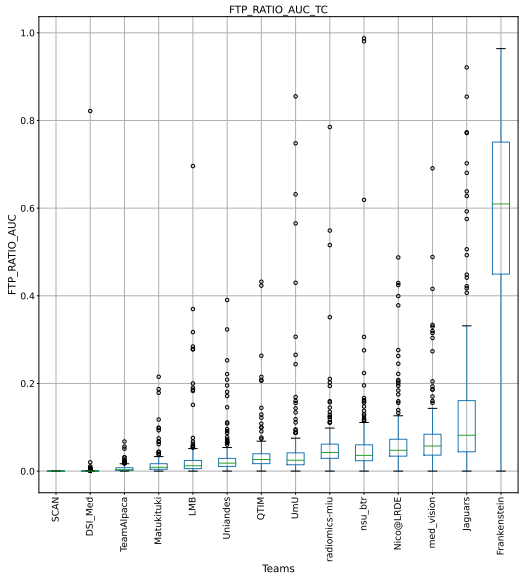

Figure 19: QU-BraTS 2020 boxplots depicting FTP_RATIO_AUC distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (lower is better).

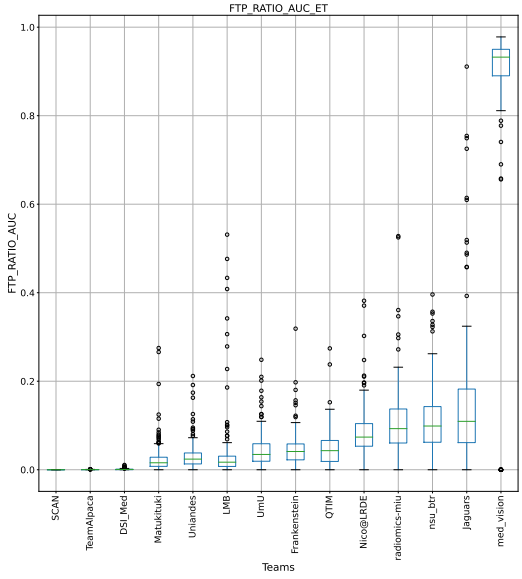

Figure 20: QU-BraTS 2020 boxplots depicting FTP_RATIO_AUC distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (lower is better).

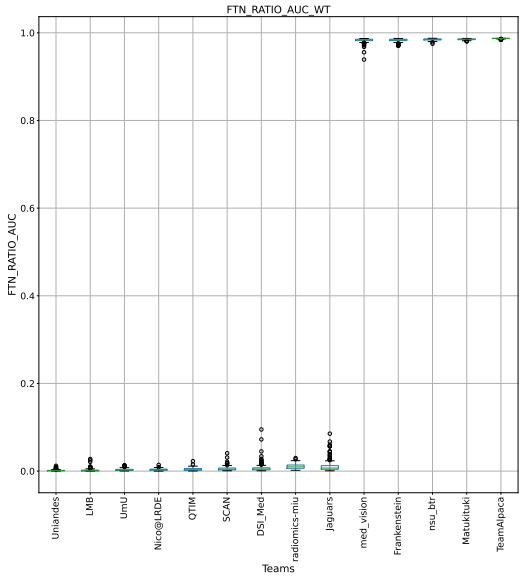

Figure 21: QU-BraTS 2020 boxplots depicting FTN_RATIO_AUC distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (lower is better).

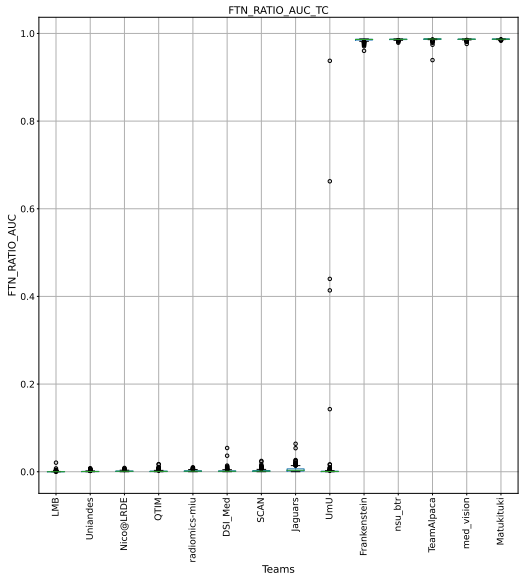

Figure 22: QU-BraTS 2020 boxplots depicting FTN_RATIO_AUC distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (lower is better).

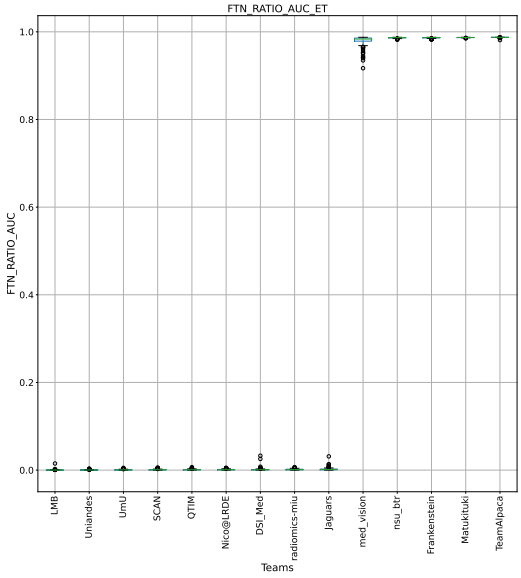

Figure 23: QU-BraTS 2020 boxplots depicting FTN_RATIO_AUC distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (lower is better).

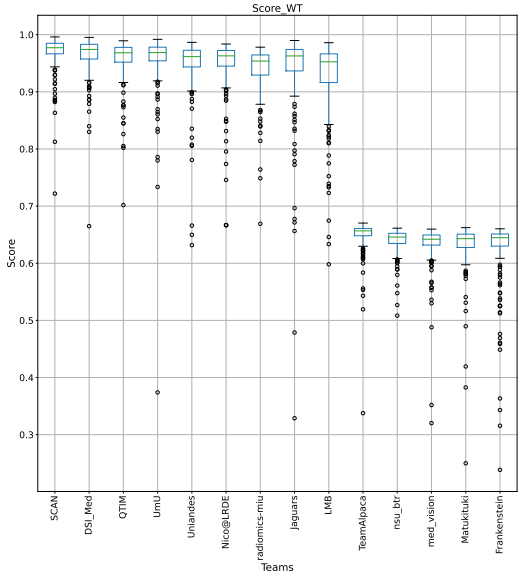

Figure 24: QU-BraTS 2020 boxplots depicting Score distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (higher is better).

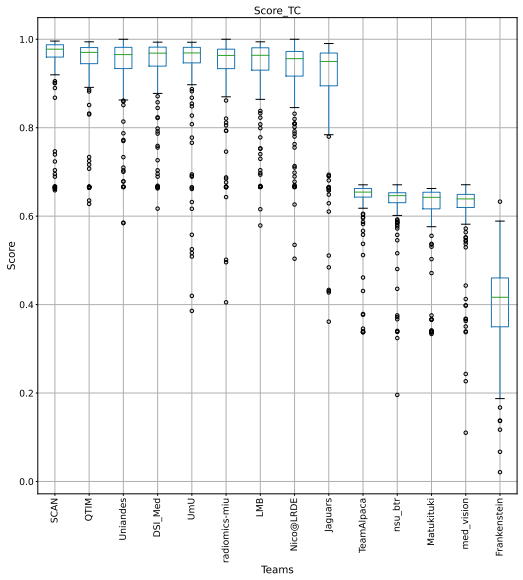

Figure 25: QU-BraTS 2020 boxplots depicting Score distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (higher is better).

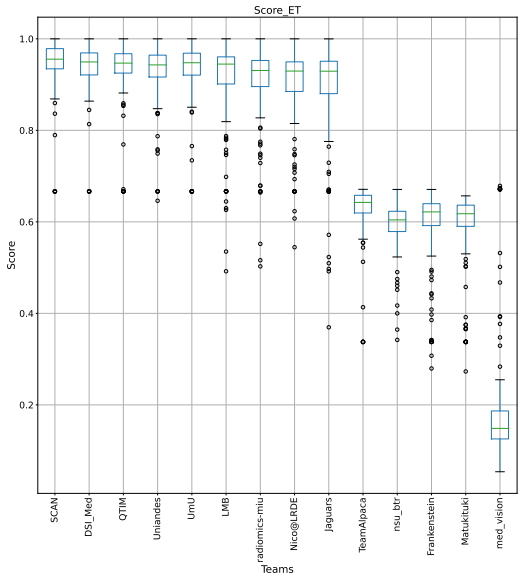

Figure 26: QU-BraTS 2020 boxplots depicting Score distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (higher is better).

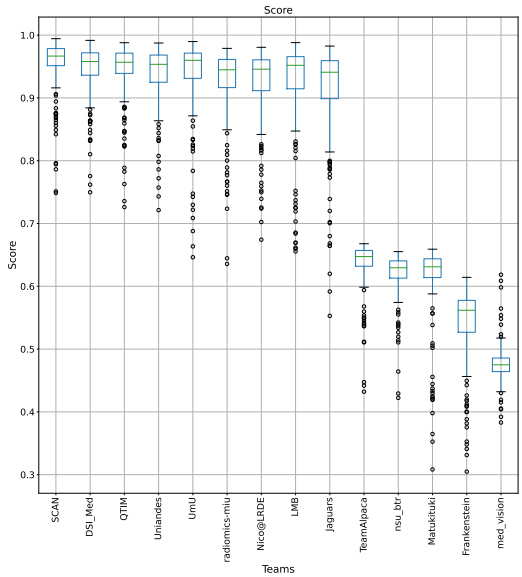

Figure 27: QU-BraTS 2020 boxplots depicting overall Score distribution for all teams across different participants on the BraTS 2020 test set (higher is better).

Team	#cases	DICE_AUC			FTP_RATIO_AUC			FTN_RATIO_AUC			Score			overall
		WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	Score
SCAN (McKinley et al., 2019)	166	0.8837	0.8253	0.8209	0.0358	0.0771	0.14958	0.01919	0.0076	0.0060	0.9429	0.9135	0.8885	0.9150
RADIOMICS-MIU (Banerjee et al., 2019a)	166	0.8595	0.8122	0.7759	0.0421	0.0906	0.12009	0.00380	0.0012	0.0008	0.9379	0.9068	0.8850	0.9099
UmU (Vu et al., 2019)	166	0.8520	0.8077	0.7892	0.0602	0.1229	0.14089	0.00334	0.0150	0.0010	0.9295	0.8899	0.8824	0.9006
xuefeng (Feng et al., 2019)	166	0.8746	0.8432	0.8120	0.0894	0.1642	0.27216	0.00969	0.0049	0.0024	0.9252	0.8914	0.8458	0.8874
UTintelligence (Amian and Soltaninejad, 2019)	162	0.7800	0.6787	0.6688	0.0117	0.0528	0.12901	0.00000	0.0000	0.0000	0.9228	0.8753	0.8466	0.8816
NVDLMED (Myronenko and Hatamizadeh, 2019)	166	0.8651	0.8203	0.8251	0.0213	0.0679	0.10958	0.49326	0.3883	0.2701	0.7835	0.7881	0.8151	0.7956
FightGliomas	166	0.8275	0.7783	0.4583	0.3172	0.2312	0.51028	0.00239	0.0008	0.0007	0.8360	0.8488	0.6491	0.7779
NIC-VICOROB	166	0.3077	0.6883	0.6393	0.5380	0.0458	0.08012	0.00000	0.0000	0.0000	0.5899	0.8808	0.8531	0.7746
LRDE_2 (Boutry et al., 2019)	166	0.8851	0.8387	0.7725	0.5930	0.7017	0.26159	0.05312	0.0439	0.0196	0.7463	0.6977	0.8304	0.7581
LRDE_VGG (Boutry et al., 2019)	166	0.8810	0.7883	0.6303	0.4930	0.7313	0.83645	0.04460	0.0280	0.0185	0.7812	0.6764	0.5918	0.6831
ANSIR	166	0.8727	0.8551	0.8349	0.0124	0.0765	0.11249	0.92500	0.9250	0.9250	0.6451	0.6179	0.5992	0.6207
med_vision (Pei et al., 2019)	166	0.8794	0.8512	0.8491	0.0203	0.0768	0.13209	0.92435	0.9253	0.9257	0.6449	0.6164	0.5971	0.6195
TEAM_ALPACA (Murugesan et al., 2019)	166	0.8768	0.8377	0.8116	0.0191	0.0707	0.10695	0.91639	0.9170	0.9228	0.6471	0.6167	0.5940	0.6192
ODU_vision_lab	166	0.8789	0.8517	0.8481	0.0212	0.0776	0.13283	0.92444	0.9253	0.9257	0.6444	0.6162	0.5965	0.6191
DRAG (Baid et al., 2019)	161	0.8890	0.8518	0.8105	0.0726	0.1312	0.13792	0.92280	0.9241	0.9243	0.6312	0.5989	0.5828	0.6043

Table 3: Final performance on the BraTS 2019 testing dataset for teams participating in the preliminary challenge on quantification of uncertainty in brain tumor segmentation task. Here, mean values for each score across all patient in the testing dataset is listed.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021.

Mehdi Amian and Mohammadreza Soltaninejad. Multi-resolution 3d CNN for MRI brain tumor segmentation and survival prediction. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I*, volume 11992 of *Lecture Notes in Computer Science*, pages 221–230. Springer, 2019. doi: 10.1007/978-3-030-46640-4_21. URL https://doi.org/10.1007/978-3-030-46640-4_21.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

Teresa Araújo, Guilherme Aresta, Luís Mendonça, Susana Penas, Carolina Maia, Ângela Carneiro, Ana Maria Mendonça, and Aurélio Campilho. Dr— graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715, 2020.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

Ujjwal Baid, Nisarg A. Shah, and Sanjay N. Talbar. Brain tumor segmentation with cascaded deep convolutional neural network. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II*, volume 11993 of *Lecture Notes in Computer Science*, pages 90–98. Springer, 2019. doi: 10.1007/978-3-030-46643-5_9. URL https://doi.org/10.1007/978-3-030-46643-5_9.

Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. *The cancer imaging archive*, 286, 2017a.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels

and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017b.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017c.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin 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*, 2018.

Laura Mora Ballestar and Veronica Vilaplana. MRI brain tumor segmentation and uncertainty estimation using 3d-UNet architectures. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 376–390. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72084-1_34. URL https://doi.org/10.1007/978-3-030-72084-1_34.

Subhashis Banerjee, Harkirat Singh Arora, and Sushmita Mitra. Ensemble of cnns for segmentation of glioma sub-regions with survival prediction. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II*, volume 11993 of *Lecture Notes in Computer Science*, pages 37–49. Springer, 2019a. doi: 10.1007/978-3-030-46643-5_4. URL https://doi.org/10.1007/978-3-030-46643-5_4.

Subhashis Banerjee, Sushmita Mitra, and B. Uma Shankar. Multi-planar spatial-ConvNet for segmentation and survival prediction in brain cancer. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 94–104. Springer International Publishing, 2019b. doi: 10.1007/978-3-030-11726-9_9. URL https://doi.org/10.1007/978-3-030-11726-9_9.

Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing uncertainty in medical image segmentation. In *Lecture Notes in Computer Science*, pages 119–127. Springer International Publishing, 2019. doi: 10.1007/978-3-030-32245-8_14. URL https://doi.org/10.1007/978-3-030-32245-8_14.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Cheng Bian, Chenglang Yuan, Jiexiang Wang, Meng Li, Xin Yang, Shuang Yu, Kai Ma, Jin Yuan, and Yefeng Zheng. Uncertainty-aware domain alignment for anatomical structure segmentation. *Medical Image Analysis*, 64:101732, 2020.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

Nicolas Boutry, Joseph Chazalon, Élodie Puybareau, Guillaume Tochon, Hugues Talbot, and Thierry Géraud. Using separated inputs for multimodal brain tumor segmentation with 3d u-net-like architectures. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I*, volume 11992 of *Lecture Notes in Computer Science*, pages 187–199. Springer, 2019. doi: 10.1007/978-3-030-46640-4_18. URL https://doi.org/10.1007/978-3-030-46640-4_18.

Robin Camarasa, Daniel Bos, Jeroen Hendrikse, Paul H. J. Nederkoorn, M. Eline Kooi, Aad van der Lugt, and Marleen de Bruijne. A quantitative comparison of epistemic uncertainty maps applied to multi-class segmentation. *CoRR*, abs/2109.10702, 2021. URL https://arxiv.org/abs/2109.10702.

Chun-Fu (Richard) Chen, Quanfu Fan, Neil Mallinar, Tom Sercu, and Rogério Schmidt Feris. Big-little net: An efficient multi-scale feature representation for visual and speech recognition. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJMHpjC9Ym.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Chengliang Dai, Shuo Wang, Hadrien Raynaud, Yuanhan Mo, Elsa Angelini, Yike Guo, and Wenjia Bai. Self-training for brain tumour segmentation with uncertainty estima-

tion and biophysics-guided survival prediction. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 514–523. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72084-1_46. URL https://doi.org/10.1007/978-3-030-72084-1_46.

Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

Christos Davatzikos, Saima Rathore, Spyridon Bakas, Sarthak Pati, Mark Bergman, Ratheesh Kalarot, Patmaa Sridharan, Aimilia Gastounioti, Nariman Jahani, Eric Cohen, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of medical imaging*, 5 (1):011018, 2018.

Laura Daza, Catalina Gómez, and Pablo Arbeláez. Cerberus: A multi-headed network for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 342–351. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72087-2_30. URL https://doi.org/10.1007/978-3-030-72087-2_30.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Jimit Doshi, Güray Erus, Mohamad Habes, and Christos Davatzikos. Deepmrseg: A convolutional deep neural network for anatomy and abnormality segmentation on MR images. *CoRR*, abs/1907.02110, 2019. URL http://arxiv.org/abs/1907.02110.

Chase Duncan, Francis Roxas, Neel Jani, Jane Maksimovic, Matthew Bramlet, Brad Sutton, and Sanmi Koyejo. Some new tricks for deep glioma segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 320–330. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72087-2_28. URL https://doi.org/10.1007/978-3-030-72087-2_28.

Xue Feng, Quan Dou, Nicholas J. Tustison, and Craig H. Meyer. Brain tumor segmentation with uncertainty estimation and overall survival prediction. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I*, volume 11992 of *Lecture Notes in Computer Science*, pages 304–314. Springer, 2019. doi: 10.1007/978-3-030-46640-4_29. URL https://doi.org/10.1007/978-3-030-46640-4_29.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017.

Florin C Ghesu, Bogdan Georgescu, Eli Gibson, Sebastian Guendel, Mannudeep K Kalra, Ramandeep Singh, Subba R Digumarthy, Sasa Grbic, and Dorin Comaniciu. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 676–684. Springer, 2019.

Vladimir Groza, Bair Tuchinov, Evgeniya Amelina, Evgeniy Pavlovskiy, Nikolay Tolstokulakov, Mikhail Amelin, Sergey Golushko, and Andrey Letyagin. Brain tumor segmentation and associated uncertainty evaluation using multi-sequences MRI mixture data preprocessing. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 148–157. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72087-2_13. URL https://doi.org/10.1007/978-3-030-72087-2_13.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1031–1040, 2020.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):2011–2023, 2020. doi: 10.1109/TPAMI.2019.2913372. URL https://doi.org/10.1109/TPAMI.2019.2913372.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.243. URL https://doi.org/10.1109/CVPR.2017.243.

Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. No new-net. In *International MICCAI Brainlesion Workshop*, pages 234–244. Springer, 2018.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14: 282, 2020.

Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer, 2016.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

Davood Karimi, Qi Zeng, Prateek Mathur, Apeksha Avinash, Sara Mahdavi, Ingrid Spadinger, Purang Abolmaesumi, and Septimiu E Salcudean. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Medical image analysis*, 57:186–196, 2019.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? NIPS'17, page 5580–5590, 2017. ISBN 9781510860964.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6965–6975, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/473447ac58e1cd7e96172575f48dca3b-Abstract.html`.

Tahsin Kurc, Spyridon Bakas, Xuhua Ren, Aditya Bagari, Alexandre Momeni, Yue Huang, Lichi Zhang, Ashish Kumar, Marc Thibault, Qi Qi, et al. Segmentation and classification in digital pathology for glioma research: challenges and deep learning approaches. *Frontiers in neuroscience*, 14:27, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V.

Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, February 2020. doi: 10.1109/tpami.2018.2858826. URL https://doi.org/10.1109/tpami.2018.2858826.

David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, HK Ng, Stefan M Pfister, Guido Reifenberger, et al. The 2021 who classification of tumors of the central nervous system: a summary. *Neuro-oncology*, 23(8):1231–1251, 2021.

Oskar Maier, Bjoern H Menze, Janina von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017.

Hugh McHugh, Gonzalo Maso Talou, and Alan Wang. 2d dense-UNet: A clinically valid approach to automated glioma segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 69–80. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72087-2_7. URL https://doi.org/10.1007/978-3-030-72087-2_7.

Richard McKinley, Michael Rebsamen, Raphael Meier, and Roland Wiest. Triplanar ensemble of 3d-to-2d cnns with label-uncertainty for brain tumor segmentation. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I*, volume 11992 of *Lecture Notes in Computer Science*, pages 379–387. Springer, 2019. doi: 10.1007/978-3-030-46640-4_36. URL https://doi.org/10.1007/978-3-030-46640-4_36.

Richard McKinley, Micheal Rebsamen, Katrin Dätwyler, Raphael Meier, Piotr Radojewski, and Roland Wiest. Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 401–411. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72084-1_36. URL https://doi.org/10.1007/978-3-030-72084-1_36.

Raghav Mehta and Tal Arbel. 3d u-net for brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 254–266. Springer, 2018.

Raghav Mehta, Angelos Filos, Yarin Gal, and Tal Arbel. Uncertainty evaluation metric for brain tumour segmentation. *arXiv preprint arXiv:2005.14262*, 2020.

Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, and Anton Becker. Qubiq - grand challenge, 2020. URL https://qubiq.grand-challenge.org/.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, October 2015. doi: 10.1109/tmi.2014.2377694. URL https://doi.org/10.1109/tmi.2014.2377694.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/95f8d9901ca8878e291552f001f67692-Abstract.html.

Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.

Gowtham Krishnan Murugesan, Sahil S. Nalawade, Chandan Ganesh Bangalore Yogananda, Benjamin C. Wagner, Fang F. Yu, Baowei Fei, Ananth J. Madhuranthakam, and Joseph A. Maldjian. Multidimensional and multiresolution ensemble networks for brain tumor segmentation. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II*, volume 11993 of *Lecture Notes in Computer Science*, pages 148–157. Springer, 2019. doi: 10.1007/978-3-030-46643-5_14. URL https://doi.org/10.1007/978-3-030-46643-5_14.

Gowtham Krishnan Murugesan, Sahil Nalawade, Chandan Ganesh, Ben Wagner, Fang F. Yu, Baowei Fei, Ananth J. Madhuranthakam, and Joseph A. Maldjian. Multidimensional

and multiresolution ensemble networks for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 448–457. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72084-1_40. URL https://doi.org/10.1007/978-3-030-72084-1_40.

Andriy Myronenko and Ali Hatamizadeh. Robust semantic segmentation of brain tumor regions from 3d mris. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II*, volume 11993 of *Lecture Notes in Computer Science*, pages 82–89. Springer, 2019. doi: 10.1007/978-3-030-46643-5_8. URL https://doi.org/10.1007/978-3-030-46643-5_8.

Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020.

Jay Patel, Ken Chang, Katharina Hoebel, Mishka Gidwani, Nishanth Arun, Sharut Gupta, Mehak Aggarwal, Praveer Singh, Bruce R. Rosen, Elizabeth R. Gerstner, and Jayashree Kalpathy-Cramer. Segmentation, survival prediction, and uncertainty estimation of gliomas from multimodal 3d MRI using selective kernel networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 228–240. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72087-2_20. URL https://doi.org/10.1007/978-3-030-72087-2_20.

Sarthak Pati, Ujjwal Baid, Maximilian Zenk, Brandon Edwards, Micah Sheller, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Jason Martin, Shadi Albarqouni, et al. The federated tumor segmentation (fets) challenge. *arXiv preprint arXiv:2105.05874*, 2021.

Linmin Pei, Lasitha Vidyaratne, Wei-Wen Hsu, Md Monibor Rahman, and Khan M. Iftekharuddin. Brain tumor classification using 3d convolutional neural network. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II*, volume 11993 of *Lecture Notes in Computer Science*, pages 335–342. Springer, 2019. doi: 10.1007/978-3-030-46643-5_33. URL https://doi.org/10.1007/978-3-030-46643-5_33.

Linmin Pei, A. K. Murat, and Rivka Colen. Multimodal brain tumor segmentation and survival prediction using a 3d self-ensemble ResUNet. In *Brainlesion: Glioma, Multiple*

Sclerosis, Stroke and Traumatic Brain Injuries, pages 367–375. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72084-1_33. URL https://doi.org/10.1007/978-3-030-72084-1_33.

Annika Reinke, Matthias Eisenmann, Minu Dietlinde Tizabi, Carole H. Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, M. Jorge Cardoso, Veronika Cheplygina, Keyvan Farahani, Ben Glocker, Doreen Heckmann-Nötzel, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Jens Kleesiek, Tahsin M. Kurç, Michal Kozubek, Bennett A. Landman, Geert Litjens, Klaus H. Maier-Hein, Bjoern H. Menze, Henning Müller, Jens Petersen, Mauricio Reyes, Nicola Rieke, Bram Stieltjes, Ronald M. Summers, Sotirios A. Tsaftaris, Bram van Ginneken, Annette Kopp-Schneider, Paul Jäger, and Lena Maier-Hein. Common limitations of image processing metrics: A picture story. *CoRR*, abs/2104.05642, 2021. URL https://arxiv.org/abs/2104.05642.

Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5): 798–819, 2010.

S. Rosas-González, I. Zemmoura, and C. Tauber. 3d brain tumor segmentation and survival prediction using ensembles of convolutional neural networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 241–254. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72087-2_21. URL https://doi.org/10.1007/978-3-030-72087-2_21.

Richard Shaw, Carole H Sudre, Sebastien Ourselin, M Jorge Cardoso, and Hugh G Pemberton. A decoupled uncertainty model for mri segmentation quality estimation. *arXiv preprint arXiv:2109.02413*, 2021.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

John Skinner, David Hall, Haoyang Zhang, Feras Dayoub, and Niko Sünderhauf. The probabilistic object detection challenge. *CoRR*, abs/1903.07840, 2019. URL http://arxiv.org/abs/1903.07840.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Yue Sun, Kun Gao, Zhengwang Wu, Guannan Li, Xiaopeng Zong, Zhihao Lei, Ying Wei, Jun Ma, Xiaoping Yang, Xue Feng, et al. Multi-site infant brain segmentation algorithms: The iseg-2019 challenge. *IEEE Transactions on Medical Imaging*, 40(5):1363–1376, 2021.

Niko Sünderhauf, Feras Dayoub, David Hall, John Skinner, Haoyang Zhang, Gustavo Carneiro, and Peter Corke. A probabilistic challenge for object detection. *Nature Machine Intelligence*, 1(9):443–443, September 2019. doi: 10.1038/s42256-019-0094-4. URL https://doi.org/10.1038/s42256-019-0094-4.

Mickael Tardy, Bruno Scheffer, and Diana Mateus. Uncertainty measurements for the reliable classification of mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 495–503. Springer, 2019.

Siddhesh Thakur, Jimit Doshi, Sarthak Pati, Saima Rathore, Chiharu Sako, Michel Bilello, Sung Min Ha, Gaurav Shukla, Adam Flanders, Aikaterini Kotrotsou, et al. Brain extraction on mri scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *NeuroImage*, 220: 117081, 2020.

Saverio Vadacchino, Raghav Mehta, Nazanin Mohammadi Sepahvand, Brennan Nichyporuk, James J. Clark, and Tal Arbel. Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. In Mattias P. Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schlaefer, and Floris Ernst, editors, *Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany*, volume 143 of *Proceedings of Machine Learning Research*, pages 787–801. PMLR, 2021. URL https://proceedings.mlr.press/v143/vadacchino21a.html.

Minh H. Vu, Tufve Nyholm, and Tommy Löfstedt. Tunet: End-to-end hierarchical brain tumor segmentation using cascaded networks. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I*, volume 11992 of *Lecture Notes in Computer Science*, pages 174–186. Springer, 2019. doi: 10.1007/978-3-030-46640-4_17. URL https://doi.org/10.1007/978-3-030-46640-4_17.

Minh H. Vu, Tufve Nyholm, and Tommy Löfstedt. Multi-decoder networks with multi-denoising inputs for tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 412–423. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72084-1_37. URL https://doi.org/10.1007/978-3-030-72084-1_37.

Guotai Wang, Wenqi Li, Maria A. Zuluaga, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans. Medical Imaging*, 37(7):1562–1573, 2018. doi: 10.1109/TMI.2018.2791721. URL https://doi.org/10.1109/TMI.2018.2791721.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.