frontiers | Frontiers in Neurology

ORIGINAL RESEARCH
published: 27 July 2022
doi: 10.3389/fneur.2022.932219

Preoperative Brain Tumor Imaging: Models and Software for Segmentation and Standardized Reporting

David Bouget[1], André Pedersen[1,2,3], Asgeir S. Jakola[4,5], Vasileios Kavouridis[6], Kyrre E. Emblem[7], Roelant S. Eijgelaar[8,9], Ivar Kommers[8,9], Hilko Ardon[10], Frederik Barkhof[11,12], Lorenzo Bello[13], Mitchel S. Berger[14], Marco Conti Nibali[13], Julia Furtner[15], Shawn Hervey-Jumper[14], Albert J. S. Idema[16], Barbara Kiesel[17], Alfred Kloet[18], Emmanuel Mandonnet[19], Domenique M. J. Müller[8,9], Pierre A. Robe[20], Marco Rossi[13], Tommaso Sciortino[13], Wimar A. Van den Brink[21], Michiel Wagemakers[22], Georg Widhalm[17], Marnix G. Witte[23], Aeilko H. Zwinderman[24], Philip C. De Witt Hamer[8,9], Ole Solheim[6,25] and Ingerid Reinertsen[1,26]*

[1] Department of Health Research, SINTEF Digital, Trondheim, Norway, [2] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway, [3] Clinic of Surgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, [4] Department of Neurosurgery, Sahlgrenska University Hospital, Gothenburg, Sweden, [5] Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, [6] Department of Neurosurgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, [7] Division of Radiology and Nuclear Medicine, Department of Physics and Computational Radiology, Oslo University Hospital, Oslo, Norway, [8] Department of Neurosurgery, Amsterdam University Medical Centers, Vrije Universiteit, Amsterdam, Netherlands, [9] Cancer Center Amsterdam, Brain Tumor Center, Amsterdam University Medical Centers, Amsterdam, Netherlands, [10] Department of Neurosurgery, Twee Steden Hospital, Tilburg, Netherlands, [11] Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Vrije Universiteit, Amsterdam, Netherlands, [12] Institutes of Neurology and Healthcare Engineering, University College London, London, United Kingdom, [13] Neurosurgical Oncology Unit, Department of Oncology and Hemato-Oncology, Humanitas Research Hospital, Università degli Studi di Milano, Milan, Italy, [14] Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA, United States, [15] Department of Biomedical Imaging and Image-Guided Therapy, Medical University Vienna, Wien, Austria, [16] Department of Neurosurgery, Northwest Clinics, Alkmaar, Netherlands, [17] Department of Neurosurgery, Medical University Vienna, Wien, Austria, [18] Department of Neurosurgery, Haaglanden Medical Center, The Hague, Netherlands, [19] Department of Neurological Surgery, Hôpital Lariboisière, Paris, France, [20] Department of Neurology and Neurosurgery, University Medical Center Utrecht, Utrecht, Netherlands, [21] Department of Neurosurgery, Isala, Zwolle, Netherlands, [22] Department of Neurosurgery, University Medical Center Groningen, University of Groningen, Groningen, Netherlands, [23] Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, Netherlands, [24] Department of Clinical Epidemiology and Biostatistics, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, Netherlands, [25] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway, [26] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway

OPEN ACCESS

Edited by:
Deqiang Qiu,
Emory University, United States

Reviewed by:
Jan Egger,
University Hospital Essen, Germany
Hamza Farooq,
University of Minnesota Twin Cities, United States

***Correspondence:**
David Bouget
david.bouget@sintef.no

Specialty section:
This article was submitted to Applied Neuroimaging, a section of the journal Frontiers in Neurology

Received: *29 April 2022*
Accepted: *23 June 2022*
Published: *27 July 2022*

Citation:
Bouget D, Pedersen A, Jakola AS, Kavouridis V, Emblem KE, Eijgelaar RS, Kommers I, Ardon H, Barkhof F, Bello L, Berger MS, Conti Nibali M, Furtner J, Hervey-Jumper S, Idema AJS, Kiesel B, Kloet A, Mandonnet E, Müller DMJ, Robe PA, Rossi M, Sciortino T, Van den Brink WA, Wagemakers M, Widhalm G, Witte MG, Zwinderman AH, De Witt Hamer PC, Solheim O and Reinertsen I (2022) Preoperative Brain Tumor Imaging: Models and Software for Segmentation and Standardized Reporting. Front. Neurol. 13:932219. doi: 10.3389/fneur.2022.932219

For patients suffering from brain tumor, prognosis estimation and treatment decisions are made by a multidisciplinary team based on a set of preoperative MR scans. Currently, the lack of standardized and automatic methods for tumor detection and generation of clinical reports, incorporating a wide range of tumor characteristics, represents a major hurdle. In this study, we investigate the most occurring brain tumor types: glioblastomas, lower grade gliomas, meningiomas, and metastases, through four cohorts of up to 4,000 patients. Tumor segmentation models were trained using the AGU-Net architecture with different preprocessing steps and protocols. Segmentation performances were assessed in-depth using a wide-range of voxel and patient-wise metrics covering

volume, distance, and probabilistic aspects. Finally, two software solutions have been developed, enabling an easy use of the trained models and standardized generation of clinical reports: Raidionics and Raidionics-Slicer. Segmentation performances were quite homogeneous across the four different brain tumor types, with an average true positive Dice ranging between 80 and 90%, patient-wise recall between 88 and 98%, and patient-wise precision around 95%. In conjunction to Dice, the identified most relevant other metrics were the relative absolute volume difference, the variation of information, and the Hausdorff, Mahalanobis, and object average symmetric surface distances. With our Raidionics software, running on a desktop computer with CPU support, tumor segmentation can be performed in 16–54 s depending on the dimensions of the MRI volume. For the generation of a standardized clinical report, including the tumor segmentation and features computation, 5–15 min are necessary. All trained models have been made open-access together with the source code for both software solutions and validation metrics computation. In the future, a method to convert results from a set of metrics into a final single score would be highly desirable for easier ranking across trained models. In addition, an automatic classification of the brain tumor type would be necessary to replace manual user input. Finally, the inclusion of post-operative segmentation in both software solutions will be key for generating complete post-operative standardized clinical reports.

Keywords: metastasis, meningioma, glioma, RADS, MRI, deep learning, 3D segmentation, open-source software

1. INTRODUCTION

Prognosis in patients with brain tumors is heterogeneous with survival rates varying from weeks to several years depending on the tumor grade and type, and for which most patients will experience progressive neurological and cognitive deficit (1). Brain tumors can be classified as either primary or secondary. In the former, tumors originate from the brain itself or its supporting tissues whereas in the latter cancer cells have spread from tumors located elsewhere in the body to reach the brain (i.e., brain metastasis). According to the World Health Organization classification of tumors (2), primary brain tumors are graded by histopathological and genetic analyses and can be regrouped in 100 different subtypes with frequent to relatively rare occurrences. Among the most frequent subtypes, tumors arising from the brain's supportive cell population (i.e., glial tissue) are referred to as gliomas. The more aggressive entities are labeled as high-grade gliomas (HGGs) and are graded between 3 and 4, while the less aggressive entities are referred to as diffuse lower grade gliomas (LGGs) and are graded between 2 and 3. Tumors arising from the meninges, which form the external membranous covering the brain, are referred to as meningiomas. Aside from the aforementioned large categories, other and less frequent tumor types exist (e.g., in the pituitary, sellar, or pineal regions). Each tumor category has distinct biology, prognosis, and treatment (3, 4). The most common primary malignant brain tumor type in adults is high-grade glioma which remains among the most difficult cancers to treat with a limited 5-year overall survival (5).

For patients affected by brain tumors, prognosis estimation and treatment decisions are made by a multidisciplinary team (including neurosurgeons, oncologists, and radiologists), and based on a set of preoperative MR scans. High accuracy in the preoperative diagnostics phase is of utmost importance for patient outcomes. Judgments concerning the complexity or radicality of surgery, or the risks of postoperative complications hinge on data gleaned from MR scans. Additionally, tumor-specific characteristics such as volume and location, or cortical structures profile can to a large degree be collected (6). Retrospectively, such measurements can be gathered from the analysis of surgical cohorts, multicenter trials, or registries in order to devise patient outcome prediction models (7–9). Reliable measurements and reporting of tumor characteristics are, therefore, instrumental in patient care. Standard reporting and data systems (RADSs) have been established for several solid tumors such as prostate cancer (10) and lung cancer (11). Very few attempts have been made for brain cancer in general (12) or high-grade gliomas (13). The main goal of RADSs is to provide rules for imaging techniques, terminology of reports, definitions of tumor features, and treatment response to reduce practice variation and obtain reproducible tumor classification. A broad implementation can facilitate collaborations and stimulate evaluation for the development and improvement of RADSs.

Currently, the lack of standardized and automatic methods for tumor detection in brain MR scan represents a major hurdle toward the generation of clinical reports incorporating a wide range of tumor characteristics. Manual tumor delineation or assessment by radiologists is time-consuming and subject to intra and inter-rater variations that are difficult to characterize (14)

and, therefore, rarely done in clinical practice. As a result, informative tumor features (e.g., location or volume) are often estimated from the images solely based on crude measuring techniques (e.g., eyeballing) (15).

1.1. Related Study

From the fast-growing development in the field of deep learning, convolutional neural networks have demonstrated impressive performance in various segmentation tasks and benchmark challenges, with the added-value of being fully automatic and deterministic (16). Regarding brain tumor segmentation, performances have specifically been assessed on the Brain Tumor Segmentation Challenge (BraTS) dataset (17, 18). Occurring every year since 2012, the challenge focuses on gliomas (i.e., HGGs and LGGs) and has reached a notable cohort size with a total of 2,040 patients included in the 2021 edition, and multiple MR sequences included for each patient (i.e., T1c, T1w, T2, FLAIR). Segmentation performance has been assessed using the Dice similarity coefficient and the 95th percentile Hausdorff distance (HD95) as metrics (19). The current state-of-the-art is an extension of the nnU-Net architecture (20) with an asymmetrical number of filters between the encoding and decoding paths, the substitution of all batch normalization layers by group normalization, and the addition of axial attention (21). An average Dice score of 85% together with a 17.70 mm HD95 were obtained for the enhancing tumor segmentation task in high-grade gliomas. The segmentation of other brain tumor types has been sparsely investigated in the literature in comparison, possibly due to a lack of open-access annotated data, as illustrated by recent reviews or studies investigating brain tumor segmentation in general (22, 23). Grovik et al. used a multicentric and multi-sequence dataset of 165 metastatic patients to train a segmentation model with the DeepLabV3 architecture (24, 25). The best segmentation results were around 79% Dice score with 3.6 false positive detections per patient on average. Other prior studies have focused on using variations of the DeepMedic architecture (26), using contrast-enhanced T1-weighted MRI volumes as input, to train their segmentation models (27, 28). Datasets were of a similar magnitude with around 200 patients. However, in both cases the test sets were limited to up to 20 patients, making it difficult to assess the generalization ability of the trained models in the absence of cross-validation studies. Obtained average Dice scores over the contrast-enhancing tumor were approximating 75%, with almost 8 false positive detections per patient. From a recent review on the use of machine learning applied to different meningioma-related tasks using MRI scans (29), more than 30 previous studies have investigated automatic diagnosis or grading but only a handful focused on the segmentation task. In addition, the datasets' magnitude used for segmentation purposes has been consistently smaller than for the other tasks, with barely up to 126 patients in the reported studies. Laukamp et al. reported the best Dice scores using well-known 3D neural network architectures such as DeepMedic and BioMedIA, though at the expense of heavy preprocessing techniques the likes of atlas registration (30, 31). In a previous study, we achieved equally promising performance using an attention-based U-Net architecture, reaching an average Dice score of up to 88% on contrast-enhanced T1-weighted MRI volumes (32). In addition, the cross-validation studies performed over up to 600 patients with a wide range of tumor sizes, coming from the hospital and the outpatient clinic, exhibited a proper ability to generalize from the trained models.

To summarize, with the exception of the BraTS challenge, there is a dearth of high-quality MRI datasets for brain tumor segmentation. Furthermore, open-access pretrained models and inference codes are scarce and can be cumbersome to operate, hence hindering the generation of private datasets for brain tumor segmentation tasks. On the other hand, open-source tools are being developed to assist in image labeling and the generation of AI models for clinical evaluation, such as MONAI Label (33) or Biomedisa (34). Yet, they do not integrate nor provide access to the latest and highest performing brain tumor segmentation models from the literature, or provide only semi-automatic methods hence requiring manual inputs from the user. From a validation standpoint, the focus has been on reporting Dice scores and often Hausdorff distances, while many other meaningful and possibly more relevant metrics exist and could be investigated to better highlight the strengths and weaknesses of the different segmentation methods (35, 36).

The literature on RADSs for brain tumors is equally scarce with only few attempts for preoperative glioblastoma surgery (13) or post-treatment investigation (37). In the former, automatic segmentation and computation of relevant tumor features were provided, and an excellent agreement has been shown between characteristics computed over the manual and automatic segmentation. In the latter, the interpretation of the post-treatment MR scans was provided using a structured set of rules but deprived of any automatic tumor segmentation or image analysis support.

1.2. Contributions

While research is exceedingly ahead for glioma segmentation under the aegis of the BraTS challenge community, the segmentation of meningiomas and metastases is trailing behind. In addition, validation studies in the literature have too often been dominated by Dice score reporting and a broader inspection is essential to ensure clinical relevance. Finally, the outcome of this research is often not readily available, especially for the intended end-users who are clinicians without programming experience. As such, the contributions of our study are: (i) the training of robust segmentation models for glioblastomas, lower grade gliomas, meningiomas, and metastases assessed using a panel of more than 20 different metrics to better highlight performance, (ii) the development of two software solutions enabling easy and fully automatic use of the trained models and tumor features computation: Raidionics and Raidionics-Slicer, and (iii) open-access models and source code for the software and validation metrics computation.

2. DATA

For this study, four different datasets have been assembled, one for each main tumor type considered: glioblastoma, lower grade glioma, meningioma, and metastasis. The tumor type

TABLE 1 | Overview of the datasets gathered for the four brain tumor types considered.

Tumor type	Sequence type	# patients	# sources	Volume average (ml)	Volume range (ml)
Glioblastoma	T1c	2134	15	34.37 ± 28.83	[0.01, 243.39]
LGG	FLAIR	659	4	51.71 ± 78.60	[0.14, 478.83]
Meningioma	T1c	719	2	19.40 ± 28.62	[0.07, 209.38]
Metastasis	T1c	396	2	17.53 ± 17.97	[0.01, 114.77]

Only one MRI sequence is available for each patient, and T1c corresponds to Gd-enhanced T1-weighted MR scans.

was assessed at the time of surgery, when applicable, following the currently applicable guidelines (i.e., either (38) or (39)). Tumors were manually segmented in 3D by trained raters using as support either a region growing algorithm (40) or a grow cut algorithm (41, 42), and subsequent manual editing. Trained raters were supervised by neuroradiologists and neurosurgeons. On contrast-enhanced T1-weighted scans, the tumor was defined as gadolinium-enhancing tissue including non-enhancing enclosed necrosis or cysts. On FLAIR scans, the tumor was defined as the hyperintense region. The four datasets are introduced in-depth in the subsequent sections. An overall summary of the data available is reported in **Table 1**, and some visual examples are provided in **Figure 1**.

2.1. Glioblastomas

The glioblastoma dataset is made of a total of 2,134 Gd-enhanced T1-weighted MRI volumes originating from 14 different hospitals, and one public challenge.

The first 1,841 patients have been collected from 14 different hospitals worldwide: 38 patients from the Northwest Clinics, Alkmaar, Netherlands (ALK); 97 patients from the Amsterdam University Medical Centers, location VU medical center, Netherlands (AMS); 86 patients from the University Medical Center Groningen, Netherlands (GRO); 103 patients from the Medical Center Haaglanden, the Hague, Netherlands (HAG); 75 patients from the Humanitas Research Hospital, Milano, Italy (MIL); 74 patients from the Hôpital Lariboisière, Paris, France (PAR); 134 patients from the University of California San Francisco Medical Center, U.S. (SFR); 49 patients from the Medical Center Slotervaart, Amsterdam, Netherlands (SLO); 153 patients from the St Elisabeth Hospital, Tilburg, Netherlands (TIL); 171 patients from the University Medical Center Utrecht, Netherlands (UTR); 83 patients from the Medical University Vienna, Austria (VIE); 72 patients from the Isala hospital, Zwolle, Netherlands (ZWO); 456 patients from the St. Olavs Hospital, Trondheim University Hospital, Norway (STO); and 249 patients from the Sahlgrenska University Hospital, Gothenburg, Sweden. An in-depth description of most cohorts can be found in a recent study (13). The remaining 293 patients correspond to the training set of the BraTS challenge (edition 2020) but have already undergone preprocessing transformations such as skull-stripping.

Overall, MRI volume dimensions are covering $[159; 896] \times [86; 896] \times [17; 512]$ voxels, and the voxel size ranges $[0.26; 1.25] \times [0.26; 2.00] \times [0.47; 7.50]$ mm^3. An average MRI volume is $[303 \times 323 \times 193]$ pixels with a spacing of $[0.86 \times 0.84 \times 1.24]$ mm^3.

2.2. Lower Grade Gliomas

The lower grade glioma dataset is made of a total of 659 FLAIR MRI volumes, with mostly grade 2 diffuse gliomas, coming from four different hospitals: 330 patients from the Brigham and Women's Hospital, Boston, USA; 165 patients from the St. Olavs Hospital, Trondheim University Hospital, Norway; 154 patients from the Sahlgrenska University Hospital, Gothenburg, Sweden; and 10 from the University Hospital of North Norway, Norway.

Overall, MRI volume dimensions are covering $[192; 576] \times [240; 640] \times [16; 400]$ voxels, and the voxel size ranges $[0.34; 1.17] \times [0.34; 1.17] \times [0.50; 8.0]$ mm^3. An average MRI volume is $[349 \times 363 \times 85]$ pixels with a spacing of $[0.72 \times 0.72 \times 4.21]$ mm^3.

2.3. Meningiomas

The meningioma dataset is made of 719 Gd-enhanced T1-weighted MRI volumes, mostly built around a dataset previously introduced (43), showcasing patients either followed at the outpatient clinic or recommended for surgery at the St. Olavs Hospital, Trondheim University Hospital, Norway.

Overall, MRI volume dimensions are covering $[192; 512] \times [224; 512] \times [11; 290]$ voxels, and the voxel size ranges $[0.41; 1.05] \times [0.41; 1.05] \times [0.60; 7.00]$ mm^3. An average MRI volume is $[343 \times 350 \times 147]$ pixels with a spacing of $[0.78 \times 0.78 \times 1.67]$ mm^3.

2.4. Metastases

The metastasis dataset is made of a total of 396 Gd-enhanced T1-weighted MRI volumes, collected from two different hospitals: 329 patients from the St. Olavs Hospital, Trondheim University Hospital, Norway; and 67 patients from Oslo University Hospital, Oslo, Norway.

Overall, MRI volume dimensions are covering $[128; 560] \times [114; 560] \times [19; 561]$ voxels, and the voxel size ranges $[0.43; 1.33] \times [0.43; 1.80] \times [0.45; 7.0]$ mm^3. An average MRI volume is $[301 \times 370 \times 289]$ pixels with a spacing of $[0.85 \times 0.76 \times 1.08]$ mm^3.

3. METHODS

First, the process for automatic brain tumor segmentation including data preprocessing, neural network architecture, and training design is introduced in Section 3.1. Second, the tumor characteristics extraction process, using the generated tumor segmentation as input, is summarized in Section 3.2. Finally, a description of the two developed software solutions for performing segmentation and standardized reporting is given in Section 3.3.

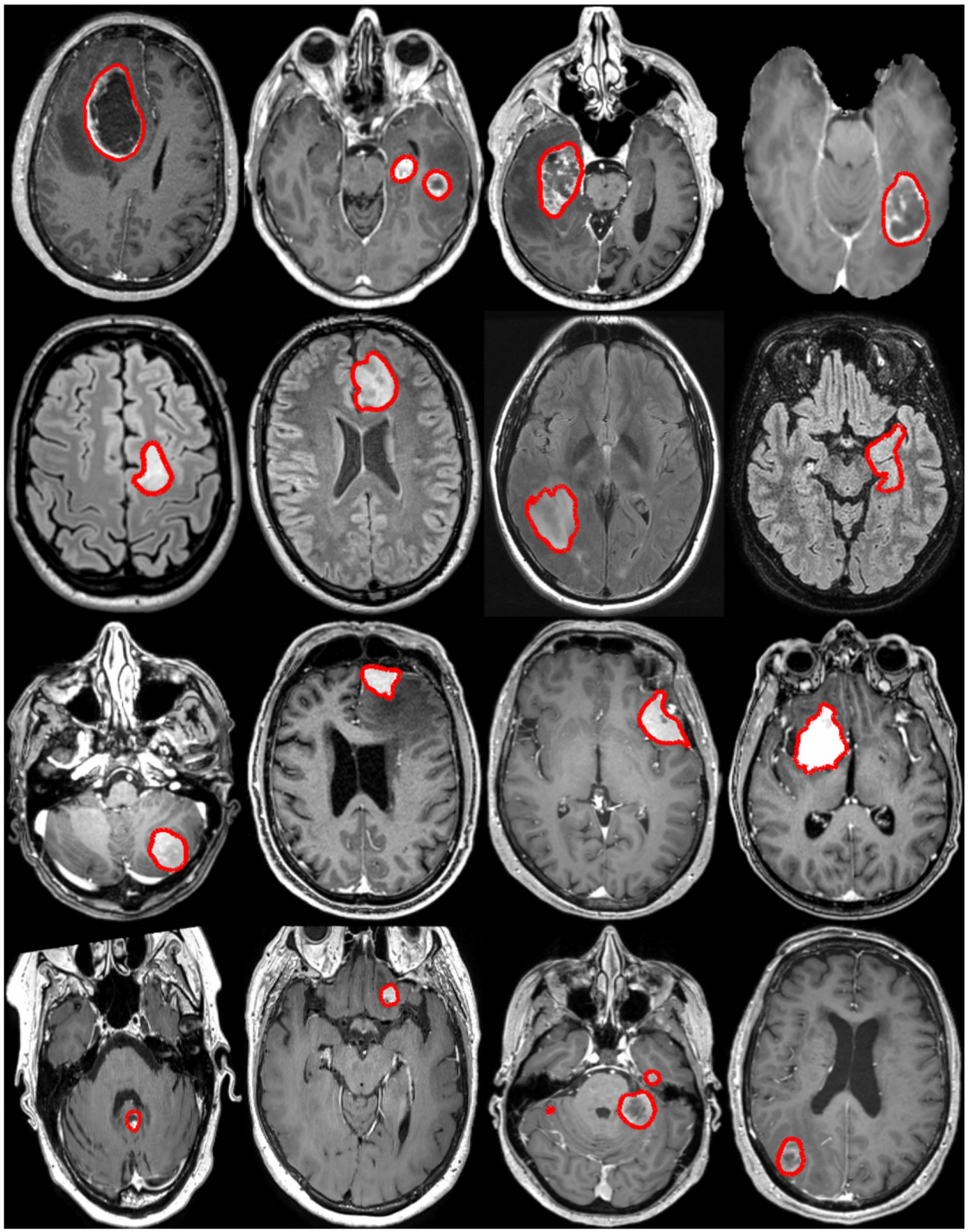

FIGURE 1 | Examples of brain tumors from the raw MRI volumes collected in this study. Each row illustrates a tumor type: glioblastoma, lower grade glioma, meningioma, and metastasis (from top to bottom). The manual annotation contours are overlaid in red.

TABLE 2 | Summary of the model training strategy followed for each tumor type.

Tumor type	Preprocessing	Strategy	Protocol
Glioblastoma	(ii) skull-stripping	(i) from-scratch	(i) leave-one-out
LGG	(i) tight clipping	(i) from-scratch	(ii) 5-fold
Meningioma	(i) tight clipping	(i) from-scratch	(ii) 5-fold
Metastasis	(ii) skull-stripping	(ii) transfer-learning	(ii) 5-fold

3.1. Tumor Segmentation

The architecture selected to train segmentation models for each brain tumor type is AGU-Net, which has shown to perform well on glioblastoma and meningioma segmentation (32, 44). In the following, the different training blocks are presented with some inner variations specified by roman numbers inside brackets. A global overview is provided in **Table 2** summarizing used variants.

Architecture: Single-stage approach leveraging multi-scale input and deep supervision to preserve details, coupled with a single attention module. The loss function used was the class-averaged Dice loss, excluding the background. The final architecture was as described in the original article with 5 levels and [16, 32, 128, 256, 256] as convolution blocks.

Preprocessing: The following preprocessing steps were used:

1. resampling to an isotropic spacing of $1\,\mathrm{mm}^3$ using spline interpolation of order 1 from NiBabel [1].
2. (i) tight clipping around the patient's head, excluding the void background, or (ii) skull-stripping using a custom brain segmentation model.
3. volume resizing to $128 \times 128 \times 144$ voxels using spline interpolation of order 1.
4. intensity normalization to the range $[0, 1]$.

Training strategy: Models were trained using the Adam optimizer over a batch size of 32 samples with accumulated gradients (actual batch size 2), stopped after 30 consecutive epochs without validation loss improvement, following either: (i) training from scratch with $1e^{-3}$ initial learning rate, or transfer learning with an initial learning rate of $1e^{-4}$ fine-tuning over the best glioblastoma model.

For the data augmentation strategy, the following transforms were applied to each input sample with a probability of 50%: horizontal and vertical flipping, random rotation in the range $[-20°, 20°]$, and translation up to 10% of the axis dimension.

Training protocol: Given the magnitude difference within our four datasets, two different protocols were considered: (i) a three-way split at the hospital level whereby MRI volumes from one hospital constituted the validation fold; MRI volumes from a second hospital constituted the test fold; and the remaining MRI volumes constituted the training fold. As such, each hospital was used in turn as the test set in order to properly assess the ability of the different models to generalize. Or (ii) a 5-fold cross-validation with a random two-way split over MRI volumes whereby four folds are used in turn as a training set and the remaining one as a validation set, without the existence of a proper separate test set.

[1]https://github.com/nipy/nibabel

3.2. Preoperative Clinical Reporting

For the generation of standardized preoperative clinical reports in a reproducible fashion, the computation of tumor characteristics was performed after alignment to a standard reference space. As described in-depth in our previous study (13), the reference space was constituted by the symmetric Montreal Neurological Institute ICBM2009a atlas (MNI) (45). The atlas space did not possess any brain average as FLAIR sequence, the T1 atlas file was used for all tumor types.

For each tumor type, the collection of features includes volume, laterality, multifocality, cortical structure location profile, and subcortical structure location profile. Specifically tailored for glioblastomas, resectability features are, therefore, not available for the other brain tumor types.

3.3. Proposed Software

In order to make our models and tumor features easily available to the community, we have developed two software solutions. The first one is a stand-alone software called Raidionics, and the second one is a plugin to 3D Slicer given its predominant and widespread use in the field (46). Both solutions provide access to a similar back-end including inference and processing code. However, the GUI and intended user interactions differ. The trained models are stored in a separate online location and are downloaded on the user's computer at runtime. Models can be improved over time and a change will be automatically detected, resulting in the replacement of outdated models to the user's machine. Raidionics can be seen as an improved solution to our initial GSI-RADS software, covering not only glioblastomas but all major brain tumor types, offering the option to compute a similar standardized report, and providing a refined graphical user interface enabling the user to visually assess the results.

3.3.1. Stand-Alone Solution: Raidionics

The software proposes two modes: (i) single-use where only one patient is to be processed and results can be visually assessed in the 2D viewer, and (ii) batch-mode whereby a collection of patients can be processed sequentially without any viewing possibility. In each mode, the option is left to the user to solely perform tumor segmentation or to compute the whole set of tumor characteristics and generate the standardized report. For each patient, the software expects an MRI scan as input (i.e., Gd-enhanced T1-weighted or FLAIR sequence) and the tumor type must be manually selected. Additionally, a pre-existing tumor segmentation mask can be provided to bypass the automatic segmentation, if collecting the tumor characteristics is the main interest and manual annotations have been performed beforehand. The total set of processed files saved on disk includes the standardized reports, brain and tumor segmentation masks in both patient and MNI space, cortical and subcortical structures masks in both patient and MNI space, and the registration files to navigate from patient to MNI space. To complement the reporting and give the possibility for follow-up statistical studies,

the complete set of computed features is also provided in comma separated value format (i.e., .csv).

The software has been developed in Python 3.6.9, using PySide2 v5.15.2 for the graphical user interface, and only uses the Central Processing Unit (CPU) for the various computations. The software has been tested and is compatible with Windows ($\geq$ 10), macOS ($\geq$ Catalina 10.15), and Ubuntu Linux ($\geq$ 18.04). An illustration of the software is provided in **Figure 2**. Cross-platform installers and source code are freely available at https://github.com/dbouget/Raidionics.

3.3.2. 3D Slicer Plugin: Raidionics-Slicer

The 3D Slicer plugin has been developed using the DeepInfer plugin as baseline (47) and is mostly intended for tumor segmentation purposes. Through a slider, the possibility is provided to manually alter the probability threshold cutoff in order to refine the proposed binary mask. Further manual editing can be performed thereafter using the existing 3D Slicer functionalities. The back-end processing code has been bundled into a Docker image for convenience, and therefore, administrator rights are required for the end-user to perform the installation locally. The same inputs, behavior, and outputs can be expected as for the stand-alone software.

The GitHub repository for the 3D Slicer plugin can be found at https://github.com/dbouget/Raidionics-Slicer, and an illustration is provided in **Figure 3**.

4. VALIDATION STUDIES

In the validation studies, only the automatic segmentation performances are assessed. The clinical validity and relevance of the extracted tumor features have been addressed thoroughly in a previous study (13). To better grasp the different aspects of the segmentation performance, a wider set of metrics is studied as described in Section 4.1. For the voxel-wise segmentation task, only two classes are considered as the whole tumor extent (including contrast-enhancing regions, cysts, and necrosis) is the target: non-tumor tissue or tumor tissue. In that sense, a positive voxel is a voxel exhibiting tumor tissue, whereas a negative voxel is a voxel exhibiting background or normal tissue.

4.1. Metrics

Following a review of metrics for evaluating 3D medical image segmentation (36), a broad spectrum of 25 metrics was selected, computed either voxel-wise or instance-wise,

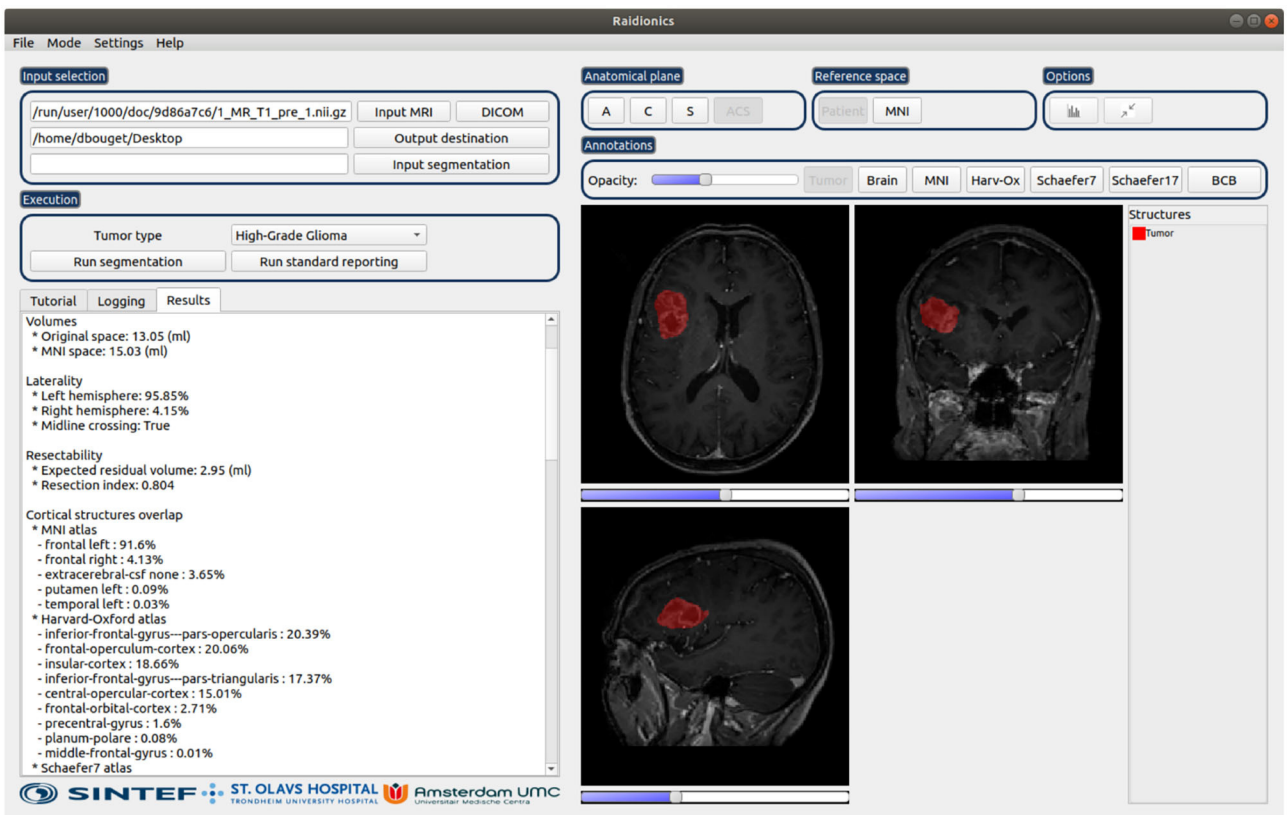

FIGURE 2 | Illustration of the Raidionics software after generating the standardized report for a patient suffering from glioblastoma. The left side presents the tumor characteristics belonging to the report, whereas the right side offers a simplistic view.

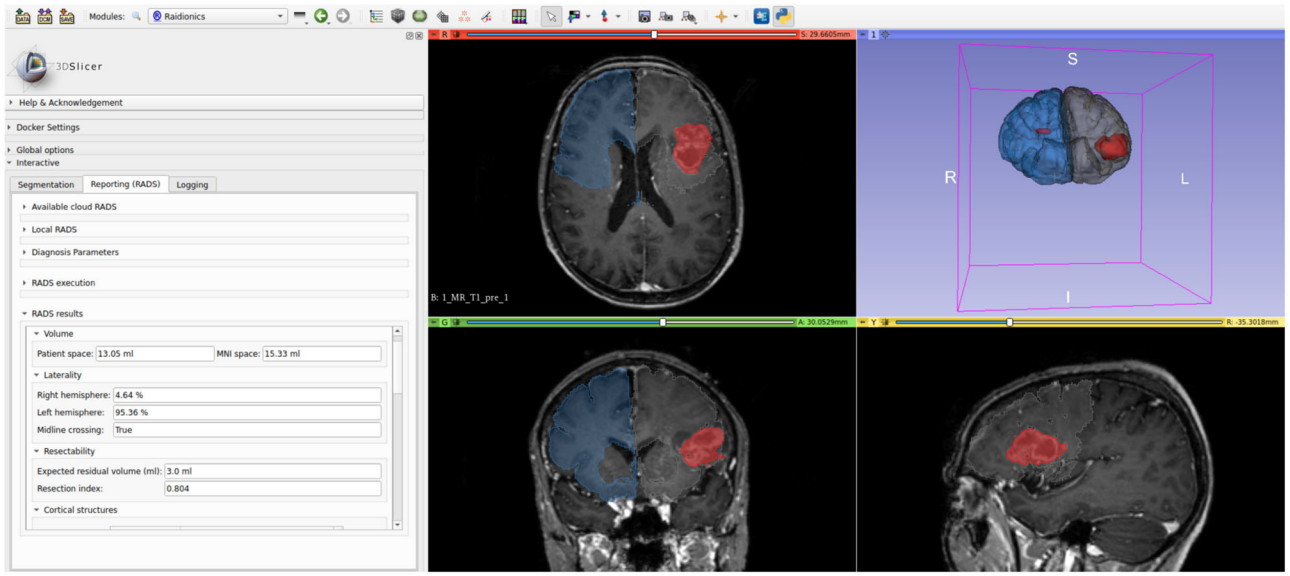

FIGURE 3 | Illustration of the Raidionics-Slicer plugin after generating the standardized report for a patient suffering from glioblastoma.

and grouped according to the following categories: overlap-based, volume-based, information theory-based, probabilistic, and spatial distance-based.

4.1.1. Voxel-Wise:

For quantifying semantic segmentation performance, we have selected the following metrics computed directly and indiscriminately over all voxels of a given patient MRI volume:

1. Overlap-based: (i) True Positive Rate (TPR), also called recall or sensitivity, is the probability that an actual positive voxel will test positive; (ii) True Negative Rate (TNR), also called specificity, is the probability that an actual negative voxel will test negative; (iii) False Positive Rate (FPR), is the probability that a false alarm will be raised (i.e., a negative voxel will test positive); (iv) False Negative Rate (FNR), also called missed rate, is the probability that a true positive voxel will test negative; (v) Positive Predictive Value (PPV), also referred to as precision, is the ratio of truly positive voxels over all voxels which tested positive; (vi) Dice score (Dice), also called the overlap index and gauging the similarity of two samples, is the most commonly used metric in validating medical volume segmentation (48); (vii) Dice True Positive score (Dice-TP) is similar to the Dice score but is only computed over the true positive predictions (i.e., when the model found the tumor); (viii) Intersection Over Union (IoU), also called the Jaccard index, measures the volume similarity as the size of the intersection divided by the size of the union of two samples (49); (ix) Global Consistency Error (GCE), defined as the error measure averaged over all voxels (50).
2. Volume-based: (i) Volumetric Similarity (VS), as the absolute volume difference divided by the sum of the compared volumes (51); (ii) Relative Absolute Volume Difference (RAVD), as the relative absolute volume difference between the joint binary objects in the two images. This is a percentage value in the range $[-1.0, \infty)$ for which a 0 denotes an ideal score.
3. Information theory-based: (i) Normalized Mutual Information (MI), normalization of the mutual information score to scale the results between 0 (no mutual information) and 1 (perfect correlation) (52); (ii) Variation Of Information (VOI), measuring the amount of information lost or gained when changing from one variable to the other, in this case, to compare clustering partitions (53).
4. Probabilistic: (i) Cohen's Kappa Score (CKS), measuring the agreement between two samples (54). The metric ranges between −1.0 and 1.0 whereby the maximum value means complete agreement, and zero or lower means chance agreement; (ii) Area Under the Curve (AUC), first presented as the measure of accuracy in the diagnostic radiology (55), further adjusted for the validation of machine learning algorithms; (iii) Volume Correlation (VC), as the linear correlation in binary object volume, measured through the Pearson product-moment correlation coefficient where the coefficient ranges $[-1., 1.]$; (iv) Matthews Correlation Coefficient (MCC), as a measure of the quality of binary and multiclass classifications, taking into account true and false positives and negatives and generally regarded as a balanced measure (56). The metric ranges between −1.0 and 1.0 whereby 1.0 represents a perfect prediction, 0.0 an average random prediction, and −1.0 an inverse prediction; (v) Probabilistic Distance (PBD), as a measure of the distance between fuzzy segmentation (57).
5. Spatial-distance-based: (i) 95th percentile Hausdorff distance (HD95), measuring the boundary delineation quality (i.e., contours). The 95% version is used to make measurements

more robust to small outliers (58); (ii) the Mahalanobis distance (MHD), measuring the correlation of all points and calculated according to the variant described for the validation of image segmentation (59); (iii) Average Symmetric Surface Distance (ASSD), as the average symmetric surface distance between the binary objects in two images.

4.1.2. Instance-Wise:

For quantifying instance detection performance, we chose the following metrics, reported in a patient-wise fashion (PW) or an object-wise fashion (OW). In the latter, and in case of multifocal tumors, each focus is considered as a separate tumor. The detection threshold has been set to 0.1% Dice to determine whether an automatic segmentation is eligible to be considered as a true detection or a false positive.

1. Overlap-based: (i) Recall, as the ratio in % of tumors properly identified; (ii) Precision, as the ratio in % of tumors incorrectly detected; (iii) F1-score (F1), measuring information retrieval as a trade-off between the recall and precision (60); (iv) False Positives Per Patient (FPPP), as the average number of incorrect detections per patient.
2. Probabilistic: (i) Adjusted Rand Index (ARI), as a similarity measure between two clusters by considering all pairs of samples and counting pairs that are assigned in the same or different clusters between the model prediction and the ground truth (61). The metric ranges from -1.0 to 1.0, whereby random segmentation has an ARI close to 0.0 and 1.0, stands for a perfect match.
3. Spatial-distance-based: (i) Object Average Symmetric Surface Distance (OASSD), as the average symmetric surface distance (ASSD) between the binary objects in two volumes.

4.2. Measurements

Pooled estimates, computed from each fold's results, are reported for each measurement (62). Overall, measurements are reported as mean and SD (indicated by $\pm$) in the tables.

Voxel-wise: For semantic segmentation performance, the Dice score is computed between the ground truth volume and a binary representation of the probability map generated by a trained model. The binary representation is computed for ten different equally-spaced probability thresholds (PT) in the range $[0, 1]$.

Instance-wise: For instance detection performance, a connected components approach coupled with a pairing strategy was employed to associate ground truth and detected tumor parts. A minimum size threshold of 50 voxels was set and objects below that limit were discarded. A detection was deemed true positive for any Dice score strictly higher than 0%.

4.3. Experiments

To validate the traied models, the following set of experiments was conducted:

1. **Overall performance study:** k-fold cross-validation studies for the different tumor types for assessing segmentation performance. For easy interpretation, only Dice scores together with patient-wise and object-wise recall, precision, and F1-score values are reported.
2. **Metrics analysis:** in-depth performance comparison using the additional metrics, and confusion matrix computation between the metrics to identify redundancy in their use.
3. **Representative models selection:** identification of one final segmentation model for each tumor type, which will be made available for use in our software solutions.
4. **Speed study:** computation of the pure inference speed and the total elapsed time required to generate predictions for a new patient, obtained with CPU support and reported in seconds. The operations required to prepare the data to be sent through the network, initialize the environment, load the trained model, and reconstruct the probability map in the referential space of the original volume are accounted for. The experiment was repeated ten consecutive times over the same MRI volume for each model, using a representative sample of each dataset in terms of dimension and spacing.

5. RESULTS

5.1. Implementation Details

Results were obtained using a computer with the following specifications: Intel Core Processor (Broadwell, no TSX, IBRS) CPU with 16 cores, 64GB of RAM, Tesla V100S (32GB) dedicated GPU and a regular hard-drive. Training and inference processes were implemented in Python 3.6 using TensorFlow v1.13.1, and the data augmentation was performed using the Imgaug Python library (63). The metrics were for the most part computed manually using the equations described in the **Supplementary Material**, or alternatively using the sklearn v0.24.2 (64) and medpy v0.4.0 (65) Python libraries. The source code used for computing the metrics and performing the validation studies is made publicly available at https://github.com/dbouget/validation_metrics_computation.

5.2. Overall Performance Study

An overall summary of brain tumor segmentation performance for all four tumor subtypes is presented in **Table 3**. Meningiomas and lower grade gliomas appear more difficult to segment given average Dice scores of 75%, compared to average Dice scores of 85% for glioblastomas and metastases. A similar trend, yet with a slightly smaller gap, can be noted for the Dice-TP scores ranging between 81 and 90% with a standard deviation of approximately 15%, indicating the quality and relative stability of the trained models. From a patient-wise perspective, those results demonstrate the difficulty of achieving good recall while keeping the precision steadily above 95%. Even though a direct comparison to the literature is impossible since different datasets have been used, obtained performance is on-par if not better than previously reported performances where Dice scores have been ranging from 75 to 85%.

Regarding the lower grade glioma tumor subtype, the diffuse nature of the tumors and less pronounced gradients over image intensities are possible explanations for the lower segmentation performance. For the meningioma category, the reason for the lower Dice-score and recall values can be attributed to the larger number of small tumors ($<$ 2 ml) compared to other subtypes. In addition, outliers have been identified in this dataset whereby a small extent of the tumors were either partly enhanced

TABLE 3 | Segmentation performance summary for each tumor type.

	Voxel-wise		Patient-wise			Object-wise		
Tumor type	**Dice**	**Dice-TP**	**F1-score**	**Recall**	**Precision**	**F1-score**	**Recall**	**Precision**
Glioblastoma	85.69 ± 16.97	87.36 ± 12.17	97.40 ± 01.01	98.08 ± 01.29	96.76 ± 01.43	89.61 ± 04.11	85.78 ± 07.95	94.19 ± 02.71
LGG	75.39 ± 25.95	81.24 ± 16.01	93.60 ± 01.74	92.86 ± 03.19	94.42 ± 01.07	81.58 ± 02.25	75.58 ± 02.41	88.70 ± 03.16
Meningioma	75.00 ± 30.52	84.81 ± 15.07	90.67 ± 01.42	88.46 ± 02.12	93.25 ± 04.76	83.85 ± 03.60	80.93 ± 04.34	87.77 ± 08.30
Metastasis	87.73 ± 18.94	90.02 ± 12.80	97.54 ± 00.76	97.46 ± 01.38	97.63 ± 00.77	88.71 ± 01.34	82.80 ± 02.38	95.60 ± 01.45

because of calcification, or non-enhancing due to intraosseous growth. For all tumor types, Dice-score distributions are reported against tumor volumes in **Figure 4** for 10 equally-sized bins. For meningiomas, four bins are necessary to group tumors with a volume of up to 4 ml while only one bin is necessary for the glioblastomas, indicating a volume distribution imbalance between the two types. The diamond-shaped points outside the boxes represent cases where the segmentation model did not perform well (cf. **Supplementary Figures S1–S4**).

While tumor volumes and outlier MR scans are reasons for the discrepancy in Dice and recall values across the board, precision is rather unaffected and more stable. The nature of the convolutional neural network architecture and training strategy used can explain those results. By leveraging volumes covering the full brain, global relationships can be learned by the trained model hence reducing the confusion between tumor regions and other contrast-enhancing structures such as blood vessels. Given GPU memory limitation, the preprocessed MR scans have undergone significant downsampling, and as such small tumors are reduced to very few voxels, impacting mainly recall performance.

Finally, an average decrease of $\sim$ 10% can be noticed between patient-wise and object-wise detection metrics, whereby satellite tumors are on average an order of magnitude smaller than the main tumor, and are hence more prone to be omitted or poorly segmented by our models. Segmentation performance is illustrated in **Figure 5**. Each row corresponds to one tumor type and each column depicts a different patient.

5.3. Metrics Analysis

Side-by-side voxel-wise performances regarding the overlap-based metrics are reported in **Tables 4**, **5**. Unsurprisingly, given the good precision performance and the absence of patients without a tumor, both TNR and its opposite FPR scores are almost perfect for all tumor types. Similarly, the TPR and its opposite FNR metrics are scoring similarly to Dice. Within each tumor category, the overlap-based metrics are following the same trend whereby a higher average Dice score would correlate with a higher score for any other metrics and vice versa (e.g., IoU). An exception can be made regarding the behavior of the GCE metric, scoring on average higher for glioblastomas than for meningiomas and as such not following the same pattern as Dice. Upon careful visual inspection, the GCE metric seems to be extremely sensitive to outliers, either coming from the image quality or manual ground truth correctness (cf. top row in **Supplementary Figures S1–S4**). Given the non-normalized state of the GCE metric, and its absence of any upper bound, an extremely poor agreement between manual ground truth and automatic segmentation will result in score orders of magnitude higher than its average expression over a given dataset. Regarding the two volume-based metrics, featured rightmost in the second table, an antagonistic pattern toward Dice can be observed. The VS metric has the same cross-type trend as Dice with similar yet slightly greater scores. On the other hand, while the RAVD metric scores best over the metastasis group similar to Dice, its worst average value is obtained for the glioblastoma group, hence potentially exhibiting the same frailty toward outliers as for the GCE metric.

Next off, voxel-wise performance for information theory-based and probabilistic metrics are regrouped in **Table 6**. The MI and VOI metrics, both based on information theory, are exhibiting an inverse behavior in line with observations about the relationship between Dice and GCE metrics. The normalized mutual information metric ranges from 0.668 to 0.829 for Dice scores between 75 and 87%, showcasing stability but also correlation. On the contrary, the VOI metric expresses a behavior concurrent to GCE whereby the worst performance is obtained for the lower grade gliomas and then glioblastomas categories, while it performs best over metastases where Dice also scores the highest. Alike the aforementioned metric groups exhibiting inner discrepancies, three of the five probabilistic metrics follow a similar trend scoring high alongside Dice, with an average gap of 0.1 corresponding to a $\sim$ 10% Dice score difference. Meanwhile, the PBD metric has a behavior of its own scoring order of magnitude worse for the meningioma category than for the three other subtypes. The metric is not normalized and an extremely poor agreement between the manual ground truth and automatic segmentation would result in an extremely large score, similar to the GCE metric, hence reporting the median score, in addition, might be of interest (cf. second row in **Supplementary Figures S1–S4**).

Finally, the voxel-wise distance-based metrics are reported in **Table 7**. Similar cross-type trends can also be noted whereby the best HD95 of 4.97 mm is obtained for the glioblastoma category and the worst HD95 of 10 mm for meningiomas, heavily correlated to Dice performance. Our average HD95 results appear lower than previously reported results in the literature, however, a strong statement can hardly be made as the tumors featured can vary highly in terms of volume and number of satellites which might reflect greatly on metrics' average scores.

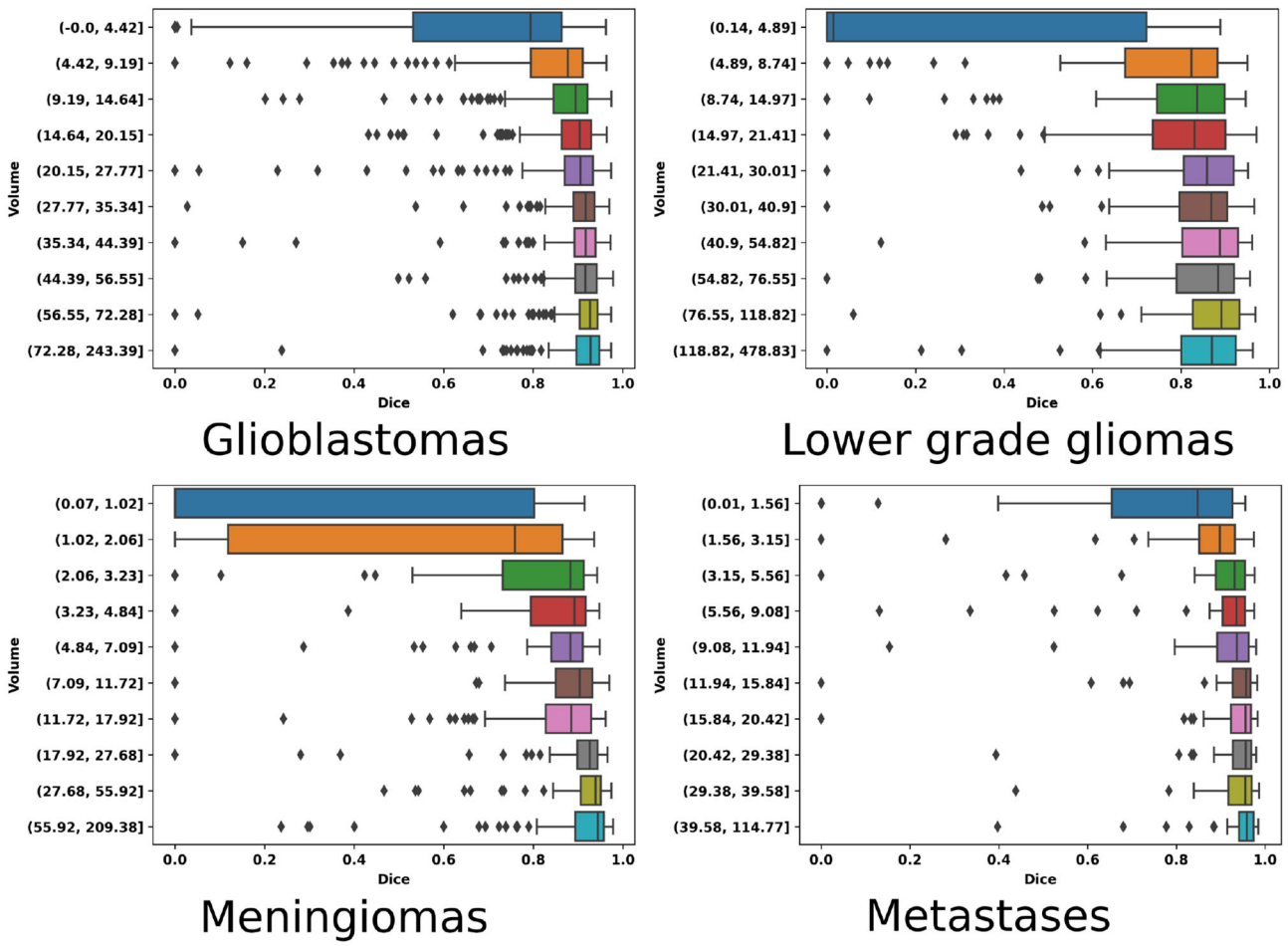

FIGURE 4 | Volume-wise (equally binned) Dice performance as boxplots for each of the four tumor types.

The other two spatial distance-based metrics display similar behavior to HD95, whereby tumor types can be ranked as follows based on best to worse performance: glioblastoma, metastasis, lower grade glioma, and meningioma.

Regarding instance-wise metrics, grouped in **Table 8**, the close OASSD average values between glioblastomas and meningiomas represent the most surprising outcome given the 5% difference in F1-score. Unsurprisingly, the lower grade glioma category achieves the highest average OASSD with 2.6 mm together with the lowest F1-score. As one might expect, the amount of FPPP correlates greatly with the average precision values obtained. Ultimately, the ARI metric generates scores extremely similar to voxel-wise Dice and correlates highly with the F1-score whereby the glioblastoma and metastasis categories obtain almost 0.1 more than for the meningioma and lower grade glioma subtypes.

For completeness, the correlation between the different metrics computed in this study has been assessed, and the results over the glioblastoma category are shown in **Table 9** (cf. other correlation matrices in **Supplementary Tables S2, S4, S6, S8**). Some metrics have been excluded given inherent correlation from their computation, such as FPR and FNR being the opposite of TNR and TPR. Similarly, metrics having computation in a voxel-wise, patient-wise, or instance-wise fashion were not considered in the matrix (i.e., recall, precision, and F1-score). Overall, the conclusions identified by analyzing the raw average results are further confirmed whereby a majority of voxel-wise metrics correlate with one another and thus do not bring any additional information to Dice. However, relevant insight can be obtained from the RAVD and GCE/VOI metrics given their low correlation to Dice and their higher sensitivity toward outliers, enabling to quantify the ability to generalize the model or potentially the quality of the data and manual ground truth (cf. third row in **Supplementary Figures S1–S4**). The correlation between HD95 and MHD appears also quite low for spatial distance-based metrics, indicating potential usefulness. Finally, in the instance-wise category, the OASSD is a stand-alone metric offering to properly assess model performance over the detection of satellite tumors. To conclude, a final pool of metrics to consider for benchmarking purposes and capturing all aspects of the segmentation performances are Dice, RAVD, VOI, HD95, MHD, and OASSD. Given the task, reporting patient-wise and

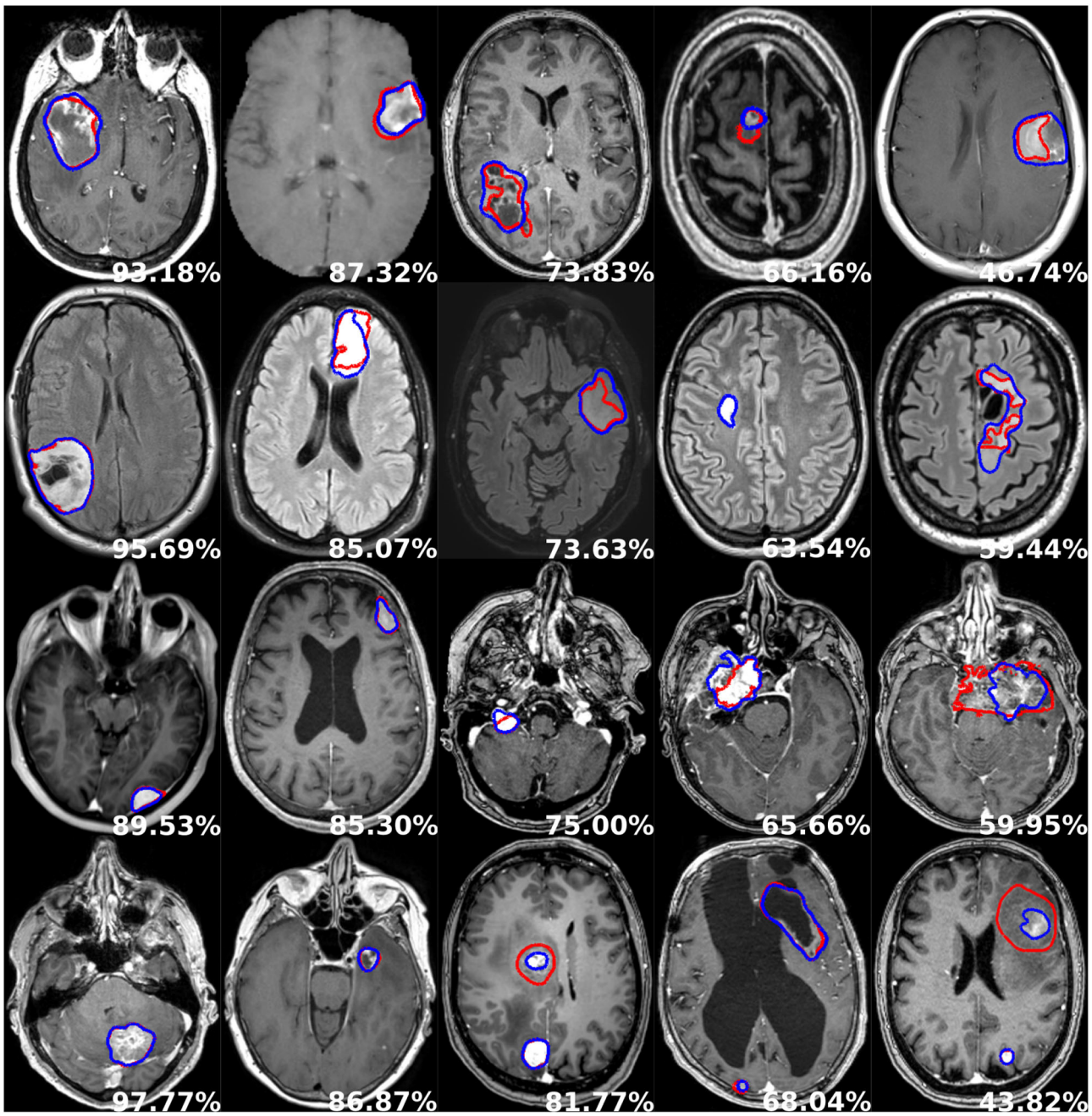

FIGURE 5 | Examples of segmentation performances. One row illustrates one tumor type: glioblastoma, lower grade glioma, meningioma, metastasis (from top to bottom), and each column depicts a different patient. The manual delineation is shown in red, the automatic segmentation in blue, and the patient-wise Dice score in white.

instance-wise recall, precision, and F1-score is always of interest because of an innate comprehension of their meaning, easy to interpret for clinicians or other end-users.

5.4. Representative Models Selection

Only one model can be provided in the software solutions for each tumor type, and the best model selection was done empirically according to the following criterion: the size of the validation or test set, average Dice score, and patient-wise F1-score performances. The exhaustive list of chosen models is the following: the model trained for fold 0 was selected for the glioblastomas, the model trained for fold 3 was selected for the lower grade gliomas, for the meningiomas the model trained for fold 2 was selected,

TABLE 4 | Voxel-wise overlap-based metrics performance summary for each tumor type.

Tumor type	TPR	TNR	FPR	FNR	PPV
Glioblastoma	87.88 ± 17.64	99.96 ± 00.06	00.04 ± 00.06	12.12 ± 17.64	87.35 ± 13.29
LGG	77.91 ± 27.89	99.90 ± 00.16	00.09 ± 00.16	22.08 ± 27.89	82.16 ± 17.01
Meningioma	77.44 ± 32.48	99.97 ± 00.04	00.02 ± 00.04	22.56 ± 32.48	84.77 ± 15.69
Metastasis	88.45 ± 20.82	99.98 ± 00.03	00.01 ± 00.03	11.54 ± 20.82	89.43 ± 16.78

TABLE 5 | Voxel-wise performance summary for each tumor type for overlap-based and volume-based metrics.

	Overlap-based				Volume-based	
Tumor type	Dice	Dice-TP	IoU	GCE (1e4)	VS	RAVD
Glioblastoma	85.69 ± 16.97	87.36 ± 12.17	77.59 ± 17.99	12.34 ± 12.57	90.43 ± 16.94	13.98 ± 171.2
LGG	75.39 ± 25.95	81.24 ± 16.01	65.72 ± 25.32	34.15 ± 46.34	82.20 ± 26.44	07.88 ± 60.14
Meningioma	75.00 ± 30.52	84.81 ± 15.07	67.13 ± 29.39	09.04 ± 17.53	80.21 ± 31.08	07.87 ± 61.31
Metastasis	87.73 ± 18.94	90.02 ± 12.80	81.56 ± 20.42	04.55 ± 07.62	91.37 ± 18.61	02.11 ± 55.35

TABLE 6 | Voxel-wise performance summary for each tumor type for information theory-based and probabilistic metrics.

	Information theory-based		Probabilistic				
Tumor type	MI	VOI	CKS	AUC	VC	MCC	PBD
Glioblastoma	0.787 ± 0.168	0.011 ± 0.009	0.856 ± 0.169	0.939 ± 0.088	0.978 ± 0.089	0.875 ± 0.122	0.840 ± 24.02
LGG	0.668 ± 0.246	0.026 ± 0.030	0.753 ± 0.259	0.889 ± 0.139	0.961 ± 0.119	0.812 ± 0.167	0.573 ± 04.82
Meningioma	0.691 ± 0.291	0.008 ± 0.013	0.749 ± 0.305	0.887 ± 0.162	0.954 ± 0.149	0.841 ± 0.171	5.358 ± 103.4
Metastasis	0.829 ± 0.191	0.004 ± 0.006	0.877 ± 0.189	0.942 ± 0.104	0.978 ± 0.100	0.901 ± 0.127	0.152 ± 0.623

TABLE 7 | Voxel-wise performance summary for each tumor type for spatial distance-based metrics.

Tumor type	HD95	MHD	ASSD
Glioblastoma	04.97 ± 09.06	00.41 ± 03.69	01.46 ± 03.22
LGG	08.37 ± 13.31	00.53 ± 03.27	02.19 ± 05.06
Meningioma	10.11 ± 21.82	00.72 ± 03.57	02.77 ± 07.91
Metastasis	07.54 ± 20.61	00.54 ± 04.56	01.73 ± 05.89

and finally for the metastases the model trained for fold 2 was selected.

5.5. Speed Study

A comparison in processing speed regarding pure tumor segmentation and complete generation of standardized reports is provided in **Table 10** when using the Raidionics software with CPU support. The high-end computer is the computer used for training the models, whereas the mid-end computer is a Windows laptop with an Intel Core Processor (*i7*@1.9GHz), and 16GB of RAM.

For the smallest MRI volumes on average, 17 s are needed to perform tumor segmentation whereas 4.5 min are required to generate the complete standardized report with the high-end computer. Unsurprisingly, the larger the MRI volume the more time required to perform the different processing operations (cf. **Supplementary Section S3**). For the largest MRI volumes overall, 54 s are needed to perform tumor segmentation whereas 15.1 min are required to generate the complete standardized report. When using the mid-end laptop, overall runtime is increased by 1.5 times for the different MRI volume sizes. On average, 9 min are necessary to generate the standardized report for MRI volumes of reasonable quality.

6. DISCUSSION

In this study, we have investigated the segmentation of a range of common main brain tumor types in 3D preoperative MR scans using a variant of the Attention U-Net architecture. We have conducted experiments to assess the performances of each trained model using close to 30 metrics and developed two software solutions for end-users to freely benefit from our segmentation models and standardized clinical reports. The main contributions are the high performances of the models, on-par with performances reported in the literature for the glioblastomas, with illustrated robustness and ability to generalize due to the multiple and widespread data sources. In addition, the two proposed open-access and open-source software solutions include our best models, together with a RADS for computing tumor characteristics. This is the first open RADS solution that supports all major brain tumor types. The software is user-friendly, requiring only a few clicks and no

TABLE 8 | Instance-wise performance for each tumor type.

Tumor type	F1-score	Recall	Precision	FPPP	ARI	OASSD
Glioblastoma	89.61 ± 04.11	85.78 ± 07.95	94.19 ± 02.71	0.078 ± 0.037	0.856 ± 0.169	01.45 ± 02.82
LGG	81.58 ± 02.25	75.57 ± 02.40	88.67 ± 03.16	0.129 ± 0.041	0.751 ± 0.259	02.60 ± 06.10
Meningioma	83.85 ± 03.60	80.93 ± 04.34	87.77 ± 08.30	0.151 ± 0.128	0.749 ± 0.305	01.62 ± 04.09
Metastasis	88.71 ± 01.34	82.79 ± 02.38	95.60 ± 01.45	0.061 ± 0.020	0.877 ± 0.189	0.672 ± 0.869

TABLE 9 | Metrics correlation matrix for glioblastoma segmentation.

	Overlap						Volume		Information theory		Probabilistic					Spatial distance			Instance-wise	
	Dice	TPR	TNR	PPV	IoU	GCE	VS	RAVD	MI	VOI	CKS	AUC	VC	MCC	PBD	HD95	MHD	ASSD	ARI	OASSD
Dice	1.0	0.7	0.29	0.62	0.98	-0.22	0.94	-0.35	0.99	-0.23	1.0	0.71	0.78	1.0	-0.34	-0.55	-0.43	-0.71	1.0	-0.3
TPR	0.7	1.0	-0.17	-0.07	0.71	-0.08	0.62	0.1	0.7	-0.08	0.7	1.0	0.51	0.71	-0.26	-0.38	-0.34	-0.47	0.7	-0.2
TNR	0.29	-0.17	1.0	0.58	0.28	-0.76	0.29	-0.36	0.33	-0.76	0.29	-0.17	0.23	0.29	-0.04	-0.16	-0.04	-0.27	0.29	-0.22
PPV	0.62	-0.07	0.58	1.0	0.64	-0.24	0.55	-0.49	0.64	-0.25	0.62	-0.07	0.47	0.63	-0.16	-0.38	-0.21	-0.47	0.62	-0.22
IoU	0.98	0.71	0.28	0.64	1.0	-0.24	0.9	-0.29	0.99	-0.24	0.98	0.71	0.71	0.99	-0.28	-0.55	-0.37	-0.7	0.98	-0.31
GCE	-0.22	-0.08	-0.76	-0.24	-0.24	1.0	-0.19	0.13	-0.3	1.0	-0.23	-0.09	-0.14	-0.23	0.02	0.18	0.03	0.29	-0.23	0.28
VS	0.94	0.62	0.29	0.55	0.9	-0.19	1.0	-0.37	0.9	-0.2	0.94	0.62	0.76	0.92	-0.36	-0.48	-0.43	-0.65	0.94	-0.26
RAVD	-0.35	0.1	-0.36	-0.49	-0.29	0.13	-0.37	1.0	-0.31	0.15	-0.35	0.1	-0.39	-0.34	0.18	0.19	0.14	0.28	-0.35	0.15
MI	0.99	0.7	0.33	0.64	0.99	-0.3	0.9	-0.31	1.0	-0.31	0.99	0.7	0.74	0.99	-0.31	-0.56	-0.4	-0.71	0.99	-0.32
VOI	-0.23	-0.08	-0.76	-0.25	-0.24	1.0	-0.2	0.15	-0.31	1.0	-0.23	-0.08	-0.15	-0.24	0.03	0.18	0.03	0.3	-0.24	0.28
CKS	1.0	0.7	0.29	0.62	0.98	-0.23	0.94	-0.35	0.99	-0.23	1.0	0.71	0.78	1.0	-0.34	-0.55	-0.43	-0.71	1.0	-0.3
AUC	0.71	1.0	-0.17	-0.07	0.71	-0.09	0.62	0.1	0.7	-0.08	0.71	1.0	0.51	0.71	-0.27	-0.38	-0.34	-0.47	0.71	-0.2
VC	0.78	0.51	0.23	0.47	0.71	-0.14	0.76	-0.39	0.74	-0.15	0.78	0.51	1.0	0.78	-0.49	-0.51	-0.58	-0.71	0.78	-0.22
MCC	1.0	0.71	0.29	0.63	0.99	-0.23	0.92	-0.34	0.99	-0.24	1.0	0.71	0.78	1.0	-0.36	-0.55	-0.44	-0.71	1.0	-0.31
PBD	-0.34	-0.26	-0.04	-0.16	-0.28	0.02	-0.36	0.18	-0.31	0.03	-0.34	-0.27	-0.49	-0.36	1.0	0.16	0.97	0.29	-0.34	0.05
HD95	-0.55	-0.38	-0.16	-0.38	-0.55	0.18	-0.48	0.19	-0.56	0.18	-0.55	-0.38	-0.51	-0.55	0.16	1.0	0.25	0.89	-0.55	0.14
MHD	-0.43	-0.34	-0.04	-0.21	-0.37	0.03	-0.43	0.14	-0.4	0.03	-0.43	-0.34	-0.58	-0.44	0.97	0.25	1.0	0.4	-0.43	0.06
ASSD	-0.71	-0.47	-0.27	-0.47	-0.7	0.29	-0.65	0.28	-0.71	0.3	-0.71	-0.47	-0.71	-0.71	0.29	0.89	0.4	1.0	-0.71	0.2
ARI	1.0	0.7	0.29	0.62	0.98	-0.23	0.94	-0.35	0.99	-0.24	1.0	0.71	0.78	1.0	-0.34	-0.55	-0.43	-0.71	1.0	-0.3
OASSD	-0.3	-0.2	-0.22	-0.22	-0.31	0.28	-0.26	0.15	-0.32	0.28	-0.3	-0.2	-0.22	-0.31	0.05	0.14	0.06	0.2	-0.3	1.0

The color intensity of each cell represents the strength of the correlation, where blue denotes direct correlation and red denotes inverse correlation.

TABLE 10 | Segmentation (Segm.) and standardized reporting (SR) execution speeds for each tumor subtype, using our Raidionics software.

		High-end computer (Desktop)		Mid-end computer (Laptop)	
	Dimensions (voxels)	Segm. (s)	SR (m)	Segm. (s)	SR (m)
LGG	394 × 394 × 80	16.69 ± 0.426	04.50 ± 0.09	28.69 ± 0.577	07.32 ± 0.07
Meningioma	256 × 256 × 170	17.21 ± 0.425	05.48 ± 0.12	31.41 ± 0.862	09.09 ± 0.32
Glioblastoma	320 × 320 × 220	21.99 ± 0.177	05.89 ± 0.03	33.65 ± 1.429	09.06 ± 0.24
Metastasis	560 × 560 × 561	59.06 ± 1.454	15.35 ± 0.41	98.54 ± 2.171	24.06 ± 0.93

programming to use, making it easily accessible for clinicians. The overall limitations are those already known for deep learning approaches whereby a higher amount of patients or data sources would improve the ability to generalize, boost segmentation performances, and increase the immunity toward rare tumor expressions. The employed architecture also struggles with smaller tumors given the large downsampling to feed the entire 3D MR scan in the network, hence the need for a better design combining local and global features either through multiple steps or ensembling.

The architecture and training strategy used in this study were identical to our previously published work considering that the intent was not to directly make advances on the segmentation task. Nevertheless, the stability and robustness

to train efficient models had been documented, alongside performance comparison to another well-known architecture [e.g., nnU-Net (20)], thus not precluding its use to train models for other brain tumor types. Aside from evident outliers in the datasets, where either tumors with partial or missing contrast uptake or suboptimal manual annotations were identified, the major pitfall of using the AGU-Net architecture lies in its struggle to segment equally satisfactorily small tumor pieces with a volume below 2 ml. Overall, the glioblastoma model is expected to be the most robust and able to generalize since patient data from 15 different sources were used. For other models trained on data from much fewer hospitals, with an expected limited variability in MR scan quality, their robustness is likely to be inferior. While larger datasets are often correlated with improved segmentation performance, the metastasis model is the best performing with the lowest amount of patients included. The relative easiness of the task from a clear demarcation of the tumor from surrounding normal tissue in contrast-enhanced T1-weighted volumes, and the potentially low variance in tumor characteristics with patient data coming from two hospitals only, can explain the results. Additionally, the metastasis model has been trained by transfer-learning using as input the second best performing glioblastoma model where the most data was used, which may have been the compelling factor. Lower-grade gliomas represent the most difficult type to manually segment since tumors are diffuse and infiltrating with an average volume in FLAIR sequences a lot higher than in T1 sequences for the other tumor types, and as such overall worse performances were expected.

The in-depth assessment of a larger pool of metrics allowed us to identify redundancy and uniqueness and proved that the Dice score is overall quite robust and indicative of expected performance. However, the sole use of the Dice score cannot cover all aspects of model performance, and spatial distance-based metrics (e.g., HD95 and MHD) are suggested to be used in conjunction as providing values uncorrelated to Dice. In addition, some metrics were identified to be more sensitive to outliers and are as such powerful to either assess the ability to generalize a model across data acquired on different scanners from multiple sources or quickly identify potential issues in a large body of data. Finally, and depending on the nature of the patients included in one's study and the number of satellite tumors, specific object-wise metrics are imperative to use (e.g., OASSD). Only a combination of various metrics computed either voxel-wise, patient-wise, or instance-wise can give the full picture of a model's performance. Unfortunately, interpreting and comparing sets of metrics can prove challenging, and as such further investigations regarding their merging into a unique informative and coherent score are fundamental [e.g., Roza (66)]. Furthermore, an inadequacy lies in the nature of the different metrics whereby some can be computed across all segmentations generated by a trained model, whereas others are exclusively eligible on true positive cases, i.e., when the model has correctly segmented to some extent of the tumor. For models generating perfect patient-wise recall, all metrics will be eligible for every segmentation. However, in this field of research and as of today, no trained model can fulfill this requirement due to the substantially large inter-patient variability. Ideally, the identification of relevant metrics, bringing unique information for interpreting the results, should not be confined to the validation studies. More metrics should be considered to be a part of the loss function computation during the training of neural network architectures. Attempts have been made toward using the Hausdorff distance as a loss function, but a direct minimization is challenging from an optimization viewpoint. For example, approximation of Hausdorff distance based on distance transforms, on morphological operations, or with circular and spherical kernels showed potential for medical image segmentation (67). In general, a careful mix between losses (e.g., Dice, cross-entropy, and HD95) is challenging to achieve and adaptive strategies might be required to avoid reaching a local minimum where overall segmentation performance may suffer (68).

As a current trend in the community, inference code and trained segmentation models are often at best available on GitHub repositories. As a consequence, only engineers, or people with some extent of knowledge in machine learning and programming, can benefit from such research advances. Besides, the research focus is heavily angled toward gliomas, due to the BraTS challenge influence, whereby segmentation models are expected to yield superior performance than for meningiomas and metastases. By developing and giving free and unrestricted access to our two proposed software solutions, we hope to facilitate more research on all brain tumor types. Willing research institutes have the opportunity to generate private annotated datasets at a faster pace than through fully manual labor by exploiting our trained models. Having made all source code available on GitHub, as customarily done, we made the effort to further make stand-alone solutions with easy-to-use GUIs. Hopefully, clinicians and other non-programming end-users should feel more comfortable manipulating such tools, available across the three major operating systems and necessitating only a computer with average hardware specifications. For the generation of standardized clinical reports, the computation of tumor characteristics relies heavily on the quality of the automatic segmentation, occasional mishaps are expected as models are not perfect and can omit the tumor. Therefore, manual inputs will be required sporadically to correct the tumor segmentation. Over time, new and better models will be generated and made available seamlessly into the two software through regular updates. For the time being, support for AGU-Net models only is provided due to its lighter codebase compared to nnU-Net, for similar overall performances. From a software bundling and deployment perspective, integrating a heavier inference framework and mixing backend engines (i.e., TensorFlow and Torch) will make it more challenging to create stable executables for Raidionics on Mac, Windows, and Ubuntu. Support for other architectures will be considered if new models clearly outperform the current models.

In the future, an approach incorporating a set of metrics and converting them into one final score would be highly

desirable (e.g., Roza). Not only would it help to automatically select the best model from a k-fold validation study from one unique score, but a proper assessment and ranking across multiple methods would be enabled. With all preoperative brain tumor types available for segmentation and reporting in our software, a key missing component is the automatic tumor type classification to supplement manual user input. Concurrently, the variety and amount of tumor characteristics to compute should be extended, considering more type-specific features similar to the resection index for glioblastomas. Alternatively, bringing a similar focus on post-operative segmentation of residual tumors is of great interest to both assess the quality of the surgery and refine the estimated patient outcome. The generation of a complete post-operative standardized clinical report would also be permitted with new features such as the extent of resection. Otherwise, intensifying the gathering of patient data from more widespread hospital centers and a larger array of MRI scanners is always of importance. The inclusion of more than one MR sequence per patient as segmentation input has the potential to boost overall performance, but at the same time might reduce models' potency as not always routinely available across all centers worldwide.

7. CONCLUSION

Efficient and robust segmentation models have been trained on pre-operative MR scans for the four main brain tumor types: glioblastoma, lower grade glioma, meningioma, and metastasis. In-depth performance assessment allowed to identify the most relevant metrics from a large panel, computed either voxel-wise, patient-wise, or instance-wise. Trained models and standardized reporting have been made publicly available and packaged into a stand-alone software and a 3D Slicer plugin to enable effortless widespread use.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because access to them is restricted under strict General Data Protection Regulation (GDPR) regulations. Requests to access the datasets should be directed to DB, david.bouget@sintef.no.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

IR, OS, PD, KE, and AJ: funding acquisition. AJ, KE, VK, IK, DB, HA, FB, LB, MB, MC, JF, SH-J, AI, BK, AK, EM, DM, PR, MR, TS, WV, MWa, GW, OS, and PD: data curation. DB, AP, IR, OS, and PD: conceptualization. DB: methodology and visualization. DB and AP: software and validation. IR, OS, and PD: supervision and project administration. DB, AP, IR, OS, AJ, KE, and PD: writing—original draft. HA, FB, LB, MB, MC, JF, SH-J, AI, BK, AK, EM, DM, PR, MR, TS, WV, MWa, GW, MWi, and AZ: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

This study was funded by the Norwegian National Advisory Unit for Ultrasound and Image-Guided Therapy (usigt.org); South-Eastern Norway Regional Health Authority; Contract Grant Nos. 2016102 and 2013069; Contract grant sponsor: Research Council of Norway; Contract Grant No. 261984; Contract grant sponsor: Norwegian Cancer Society; Contract Grant Nos. 6817564 and 3434180; Contract grant sponsor: European Research Council under the European Union's Horizon 2020 Program; Contract Grant No. 758657-ImPRESS; an unrestricted grant of Stichting Hanarth fonds, Machine learning for better neurosurgical decisions in patients with glioblastoma; a grant for public-private partnerships (Amsterdam UMC PPP-grant) sponsored by the Dutch government (Ministry of Economic Affairs) through the Rijksdienst voor Ondernemend Nederland (RVO) and Topsector Life Sciences and Health (LSH), Picturing predictions for patients with brain tumors; a grant from the Innovative Medical Devices Initiative program, project number 10-10400-96-14003; The Netherlands Organisation for Scientific Research (NWO), 2020.027; a grant from the Dutch Cancer Society, VU2014-7113; the Anita Veldman foundation, CCA2018-2-17. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

Data were processed in digital labs at HUNT Cloud, Norwegian University of Science and Technology, Trondheim, Norway.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2022.932219/full#supplementary-material

REFERENCES

1. Day J, Gillespie DC, Rooney AG, Bulbeck HJ, Zienius K, Boele F, et al. Neurocognitive deficits and neurocognitive rehabilitation in adult brain tumors. *Curr Treat Options Neurol*. (2016) 18:1–16. doi: 10.1007/s11940-016-0406-5
2. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021. WHO classification of tumors of the central nervous system: a summary. *Neuro Oncol*. (2021) 23:1231–51. doi: 10.1093/neuonc/noab106
3. DeAngelis LM. Brain tumors. *N Engl J Med*. (2001) 344:114–23. doi: 10.1056/NEJM200101113440207

4. Fisher JL, Schwartzbaum JA, Wrensch M, Wiemels JL. Epidemiology of brain tumors. *Neurol Clin*. (2007) 25:867–90. doi: 10.1016/j.ncl.2007.07.002
5. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. *Lancet*. (2018) 392:432–46. doi: 10.1016/S0140-6736(18)30990-5
6. Kickingereder P, Burth S, Wick A, Götz M, Eidel O, Schlemmer HP, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*. (2016) 280:880–9. doi: 10.1148/radiol.2016160845
7. Sawaya R, Hammoud M, Schoppa D, Hess KR, Wu SZ, Shi WM, et al. Neurosurgical outcomes in a modern series of 400 craniotomies for treatment of parenchymal tumors. *Neurosurgery*. (1998) 42:1044–55. doi: 10.1097/00006123-199805000-00054
8. Mathiesen T, Peredo I, Lönn S. Two-year survival of low-grade and high-grade glioma patients using data from the Swedish Cancer Registry. *Acta Neurochir*. (2011) 153:467–71. doi: 10.1007/s00701-010-0894-0
9. Zinn PO, Colen RR, Kasper EM, Burkhardt JK. Extent of resection and radiotherapy in GBM: A 1973 to 2007 surveillance, epidemiology and end results analysis of 21,783 patients. *Int J Oncol*. (2013) 42:929–34. doi: 10.3892/ijo.2013.1770
10. Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, et al. PI-RADS prostate imaging-reporting and data system: 2015 version 2. *Eur Urol*. (2016) 69:16–40. doi: 10.1016/j.eururo.2015.08.052
11. Dyer SC, Bartholmai BJ, Koo CW. Implications of the updated Lung CT Screening Reporting and Data System (Lung-RADS version 1.1) for lung cancer screening. *J Thorac Dis*. (2020) 12:6966. doi: 10.21037/jtd-2019-cptn-02
12. Ellingson BM, Bendszus M, Boxerman J, Barboriak D, Erickson BJ, Smits M, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro Oncol*. (2015) 17:1188–98. doi: 10.1093/neuonc/nov095
13. Kommers I, Bouget D, Pedersen A, Eijgelaar RS, Ardon H, Barkhof F, et al. Glioblastoma surgery imaging–reporting and data system: standardized reporting of tumor volume, location, and resectability based on automated segmentations. *Cancers*. (2021) 13:2854. doi: 10.3390/cancers13122854
14. Binaghi E, Pedoia V, Balbi S. Collection and fuzzy estimation of truth labels in glial tumour segmentation studies. *Comput Methods Biomech Biomed Eng*. (2016) 4:214–28. doi: 10.1080/21681163.2014.947006
15. Berntsen EM, Stensjøen AL, Langlo MS, Simonsen SQ, Christensen P, Moholdt VA, et al. Volumetric segmentation of glioblastoma progression compared to bidimensional products and clinical radiological reports. *Acta Neurochir*. (2020) 162:379–87. doi: 10.1007/s00701-019-04110-0
16. Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. (2021) 44:3523–42. doi: 10.1109/TPAMI.2021.3059968
17. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. (2014) 34:1993–2024. doi: 10.1109/TMI.2014.2377694
18. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*. (2017) 4:1–13. doi: 10.1038/sdata.2017.117
19. Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv[Preprint].arXiv:210702314*. (2021). doi: 10.48550/arXiv.2107.02314
20. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnu-net: self-adapting framework for u-net-based medical image segmentation. *arXiv[Preprint].arXiv:180910486*. (2018) doi: 10.1007/978-3-658-25326-4_7
21. Luu HM, Park SH. Extending nn-UNet for brain tumor segmentation. *arXiv[Preprint].arXiv:211204653*. (2021). doi: 10.48550/arXiv.2112.04653
22. Tiwari A, Srivastava S, Pant M. Brain tumor segmentation and classification from magnetic resonance images: review of selected methods from 2014 to 2019. *Pattern Recognit Lett*. (2020) 131:244–60. doi: 10.1016/j.patrec.2019.11.020
23. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging*. (2016) 35:1240–51. doi: 10.1109/TMI.2016.2538465
24. Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J Mag Reson Imaging*. (2020) 51:175–82. doi: 10.1002/jmri.26766
25. Grøvik E, Yi D, Iv M, Tong E, Nilsen LB, Latysheva A, et al. Handling missing MRI sequences in deep learning segmentation of brain metastases: a multicenter study. *NPJ Digit Med*. (2021) 4:1–7. doi: 10.1038/s41746-021-00398-4
26. Kamnitsas K, Ferrante E, Parisot S, Ledig C, Nori AV, Criminisi A, et al. DeepMedic for brain tumor segmentation. In: *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Athens: Springer (2016). p. 138–49.
27. Liu Y, Stojadinovic S, Hrycushko B, Wardak Z, Lau S, Lu W, et al. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. *PLoS One*. (2017) 12:e0185844. doi: 10.1371/journal.pone.0185844
28. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Comput Biol Med*. (2018) 95:43–54. doi: 10.1016/j.compbiomed.2018.02.004
29. Neromyliotis E, Kalamatianos T, Paschalis A, Komaitis S, Fountas KN, Kapsalaki EZ, et al. Machine learning in meningioma MRI: past to present. A narrative review. *J Mag Reson Imaging*. (2022) 55:48–60. doi: 10.1002/jmri.27378
30. Laukamp KR, Thiele F, Shakirin G, Zopfs D, Faymonville A, Timmer M, et al. Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric MRI. *Eur Radiol*. (2019) 29:124–32. doi: 10.1007/s00330-018-5595-8
31. Laukamp KR, Pennig L, Thiele F, Reimer R, Görtz L, Shakirin G, et al. Automated meningioma segmentation in multiparametric MRI. *Clin Neuroradiol*. (2020) 31:357–66. doi: 10.1007/s00062-020-00884-4
32. Bouget D, Pedersen A, Hosainey SAM, Solheim O, Reinertsen I. Meningioma segmentation in t1-weighted mri leveraging global context and attention mechanisms. *arXiv[Preprint].arXiv:210107715*. (2021). doi: 10.3389/fradi.2021.711514
33. Consortium TM. *Project MONAI*. Zenodo (2020).
34. Lösel PD, van de Kamp T, Jayme A, Ershov A, Faragó T, Pichler O, et al. Introducing Biomedisa as an open-source online platform for biomedical image segmentation. *Nat Commun*. (2020) 11:1–14. doi: 10.1038/s41467-020-19303-w
35. Reinke A, Eisenmann M, Tizabi MD, Sudre CH, Rädsch T, Antonelli M, et al. Common limitations of image processing metrics: a picture story. *arXiv preprint arXiv:210405642*. (2021). doi: 10.48550/arXiv.2104.05642
36. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. (2015) 15:29. doi: 10.1186/s12880-015-0068-x
37. Weinberg BD, Gore A, Shu HKG, Olson JJ, Duszak R, Voloschin AD, et al. Management-based structured reporting of posttreatment glioma response with the brain tumor reporting and data system. *J Am Coll Radiol*. (2018) 15:767–71. doi: 10.1016/j.jacr.2018.01.022
38. Fuller GN, Scheithauer BW. The 2007 Revised World Health Organization (WHO) classification of tumours of the central nervous system: newly codified entities. *Brain Pathology* (2007) 17:304–7. doi: 10.1111/j.1750-3639.2007.00084.x
39. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol*. (2016) 131:803–20. doi: 10.1007/s00401-016-1545-1
40. Huber T, Alber G, Bette S, Boeckh-Behrens T, Gempt J, Ringel F, et al. Reliability of semi-automated segmentations in glioblastoma. *Clin Neuroradiol*. (2017) 27:153–61. doi: 10.1007/s00062-015-0471-2
41. Vezhnevets V, Konouchine V. GrowCut: interactive multi-label ND image segmentation by cellular automata. In: *Proceedings of Graphicon, Vol. 1*. Novosibirsk: Citeseer (2005). p. 150–6.
42. Egger J, Kapur T, Fedorov A, Pieper S, Miller JV, Veeraraghavan H, et al. GBM volumetry using the 3D Slicer medical image computing platform. *Sci Rep*. (2013) 3:1–7. doi: 10.1038/srep01364
43. Bouget D, Pedersen A, Hosainey SAM, Vanel J, Solheim O, Reinertsen I. Fast meningioma segmentation in T1-weighted magnetic resonance imaging

volumes using a lightweight 3D deep learning architecture. *J Med Imaging*. (2021) 8:024002. doi: 10.1117/1.JMI.8.2.024002
44. Bouget D, Eijgelaar RS, Pedersen A, Kommers I, Ardon H, Barkhof F, et al. Glioblastoma Surgery Imaging-Reporting and Data System: Validation and Performance of the Automated Segmentation Task. *Cancers*. (2021) 13:4674. doi: 10.3390/cancers13184674
45. Fonov VS, Evans AC, McKinstry RC, Almli CR, Collins D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage*. (2009) 47:S102. doi: 10.1016/S1053-8119(09)70884-5
46. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Mag Reson Imaging*. (2012) 30:1323–41. doi: 10.1016/j.mri.2012.05.001
47. Mehrtash A, Pesteie M, Hetherington J, Behringer PA, Kapur T, Wells III WM, et al. DeepInfer: open-source deep learning deployment toolkit for image-guided therapy. In: *Medical Imaging 2017: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 10135*. SPIE(2017). p. 410–6.
48. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. (1945) 26:297–302. doi: 10.2307/1932409
49. Jaccard P. The distribution of the flora in the alpine zone[1]. *New Phytol*. (1912) 11:37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x
50. Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV, Vol. 2*. Vancouver, BC: IEEE (2001). p. 416–23.
51. Cárdenes R, de Luis-Garcia R, Bach-Cuadra M. A multidimensional segmentation evaluation for medical image data. *Comput Methods Programs Biomed*. (2009) 96:108–24. doi: 10.1016/j.cmpb.2009.04.009
52. Russakoff DB, Tomasi C, Rohlfing T, Maurer CR. Image similarity using mutual information of regions. In: *European Conference on Computer Vision*. Prague: Springer (2004). p. 596–607.
53. Meilă M. Comparing clusterings by the variation of information. In: *Learning Theory and Kernel Machines*. Washington: Springer (2003). p. 173–87.
54. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. (1960) 20:37–46. doi: 10.1177/001316446002000104
55. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. (1997) 30:1145–59. doi: 10.1016/S0031-3203(96)00142-2
56. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. (2000) 16:412–24. doi: 10.1093/bioinformatics/16.5.412
57. Gerig G, Jomier M, Chakos M. Valmet: a new validation tool for assessing and improving 3D object segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Utrecht: Springer (2001). p. 516–23. doi: 10.1007/3-540-45468-3_62
58. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the hausdorff distance. *IEEE Trans Pattern Anal Mach Intell*. (1993) 15:850–63. doi: 10.1109/34.232073
59. McLachlan GJ. Mahalanobis distance. *Resonance*. (1999) 4:20–6. doi: 10.1007/BF02834632
60. Chinchor N, Sundheim BM. MUC-5 evaluation metrics. In: *Fifth Message Understanding Conference (MUC-5)*: *Proceedings of a Conference Held in Baltimore, Maryland, August 25–27, 1993*. Baltimore, MD (1993).
61. Hubert L, Arabie P. Comparing partitions. *J Classificat*. (1985) 2:193–218. doi: 10.1007/BF01908075
62. Killeen PR. An alternative to null-hypothesis significance tests. *Psychol Sci*. (2005) 16:345–53. doi: 10.1111/j.0956-7976.2005.01538.x
63. Jung AB, Wada K, Crall J, Tanaka S, Graving J, Reinders C, et al. *Imgaug*. (2020). Available online at: https://github.com/aleju/imgaug (accessed on February 01, 2020).
64. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. (2011) 12:2825–30. doi: 10.48550/arXiv.1201.0490
65. Maier O, Rothberg A, Raamana PR, Bèges R, Isensee F, Ahern M, et al. *loli/medpy: MedPy 0.4.0*. Zenodo (2019).
66. Melek M, Melek N. Roza: a new and comprehensive metric for evaluating classification systems. *Comput Methods Biomech Biomed Engin*. (2021) 25:1–13. doi: 10.1080/10255842.2021.1995721
67. Karimi D, Salcudean SE. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Trans Med Imaging*. (2019) 39:499–513. doi: 10.1109/TMI.2019.2930068
68. Heydari AA, Thompson CA, Mehmood A. Softadapt: techniques for adaptive loss weighting of neural networks with multi-part loss functions. *arXiv[Preprint].arXiv:191212355*. (2019). doi: 10.48550/arXiv.1912.12355

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