Glioblastoma segmentation from early post-operative MRI: challenges and clinical impact

Ragnhild Holden Helland[1,2*][0000−0002−9592−4876], David Bouget[1][0000−0002−5669−9514], Roelant S. Eijgelaar[3,4][0000−0002−1765−4444], Philip C. De Witt Hamer[3,4][0000−0003−2988−8544], Frederik Barkhof[5,6][0000−0003−3543−3706], Ole Solheim[7,8][0000−0002−5954−4817], and Ingerid Reinertsen[1,2][0000−0003−0999−3849]

[1] Dept. Health Research, SINTEF Digital, Trondheim, Norway
[2] Dept. Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway
[3] Cancer Center Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
[4] Dept. of Neurosurgery, Amsterdam UMC, Amsterdam, The Netherlands
[5] Dept. of Radiology and Nuclear Medicine, Amsterdam UMC, Amsterdam, The Netherlands
[6] Inst. of Neurology and Healthcare Engineering, UCL, London , UK
[7] Dept. of Neurosurgery, St. Olavs hospital, Trondheim University Hospital, Trondheim, Norway
[8] Dept. of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway
*ragnhild.holden.helland@sintef.no

Abstract. Post-surgical evaluation and quantification of residual tumor tissue from magnetic resonance images (MRI) is a crucial step for treatment planning and follow-up in glioblastoma care. Segmentation of enhancing residual tumor tissue from early post-operative MRI is particularly challenging due to small and fragmented lesions, post-operative bleeding, and noise in the resection cavity. Although a lot of progress has been made on the adjacent task of pre-operative glioblastoma segmentation, more targeted methods are needed for addressing the specific challenges and detecting small lesions. In this study, a state-of-the-art architecture for pre-operative segmentation was used, trained on a large in-house multi-center dataset for early post-operative segmentation. Various pre-processing, data sampling techniques, and architecture variants were explored for improving the detection of small lesions. The models were evaluated on a dataset annotated by 8 novice and expert human raters, and the performance compared against the human inter-rater variability. Trained models' performance were shown to be on par with the performance of human expert raters. As such, automatic segmentation models have the potential to be a valuable tool in a clinical setting as an accurate and time-saving alternative, compared to the current standard manual method for residual tumor measurement after surgery.

Keywords: Segmentation · early post-operative MRI · glioblastoma.

1 Introduction

Glioblastoma is the most common primary malignant brain cancer in adults, requiring a multidisciplinary treatment approach consisting of maximum safe surgical resection followed by radiation and chemotherapy [6]. Still, the patient's prognosis remains poor with a median survival of only 12 months [18]. While extensive surgical resection is associated with longer survival [5], the significant tumor invasiveness renders a complete removal of all tumor cells unfeasible in most cases. The extent of resection (EOR), computed after surgery, is the ratio between surgically removed tumor volume and pre-operative tumor volume. Therefore, an utmost accurate EOR estimation relies on optimal segmentation of the fullest tumor extent both pre- and post-operatively. In current clinical practice, the residual tumor volume is estimated manually either through eye-balling [1] or according to the Response Assessment in Neuro-Oncology (RANO) criteria [20]. In the latter, the volume is measured as the bi-dimensional product of the largest axial diameter of the residual enhancing tumor. Exact manual post-operative volume segmentation would be favorable but is very time-consuming. In addition, this task is heavily expertise-dependent, with a high inter- and intra-rater variability [1, 19]. The MICCAI Brain Tumor Segmentation (BraTS) Challenge [14] has enabled many contributions on the task of pre-operative glioblastoma segmentation in recent years. The state-of-the-art for the task is represented by the winning challenge teams every year. Recently, the best-performing models have all been modified and ensembled versions of the U-Net [17] architecture, the nnU-Net in 2020 [11, 10], an extended version of nnU-Net in 2021 [13], an ensemble comprising nnU-Net, DeepSeg and DeepSCAN in 2022 [21], and finally an ensemble of nnU-Net and Swin UNETR trained on synthetic data in 2023 [7]. Yet, no large dataset for early post-operative segmentation is currently openly available. As such, much less progress has been made on this task. A few fully automated methods have been proposed for follow-up post-operative images [8, 12], most of which were trained on the BraTS dataset and fine-tuned on local datasets consisting of follow-up MRI scans. However, follow-up MRI scans were usually acquired from 3-12 months after surgery and not within the 72-hour time frame after surgical resection for early post-operative MRI. Hence, tumor regrowth and enhancement due to reparative changes in the tissue after surgery might be visible. A recent study presented a new dataset consisting of early post-operative and pre-operative MRI scans from 956 patient originating from 12 hospitals, with enhancing tumor tissue annotated by experts in both the pre- and post-operative scans [9]. Two top-performing architectures for pre-operative glioblastoma segmentation were trained: the nnU-Net [11] and AGU-Net [3]. Both architectures were able to segment residual tumor on an expert level using only two early post-operative MRI (EPMR) scans. The nnU-Net architecture, leveraging a patch-wise approach, had a higher pixel- and patient-wise sensitivity and achieved a superior segmentation performance compared to the AGU-Net architecture. However, nnU-Net also had a higher false positive rate and was unable to identify patients with gross total resection (i.e., without residual tumor). The AGU-Net architecture, using full MRI volumes as input, achieved a better

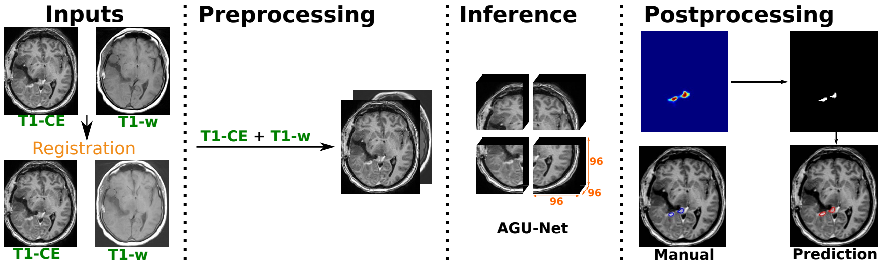

Fig. 1. Overview of the segmentation pipeline, using two early post-operative MRI scans (i.e., T1-weighted and contrast-enhanced T1-weighted. A patch-wise AGU-Net architecture with 96^3 voxels was trained.

classification of patients with gross total resection, while maintaining a similar segmentation performance as human expert raters.

In this work, the focus was brought towards improving early post-operative glioblastoma segmentation with the aim of training a model able to achieve satisfactorily segmentation and classification performance, on par with human expert raters and thus usable in a clinical setting. From the advantages of the models presented in [9], the first contribution was to investigate the particular challenges of the problem, namely the extreme class imbalance as well as the small and fragmented lesions. As a second contribution, an investigation of the impact of the ground truth quality was performed, through a qualitative analysis and interpretation of the results by comparison with several human expert raters. A deeper understanding of the challenges in early post-operative segmentation is provided, together with technical and clinical perspectives for improvement moving forward.

2 Materials and Methods

2.1 Data

A dataset comprising early post-operative MRI (EPMR) scans from 956 patients, with manual segmentations in 3D by expert neuroradiologists and neurosurgeons, was used in this study. Among these patients, 604 (65%) exhibitied an enhancing residual tumor (RT), whereas 352 (35%) were gross total resections (GTR), i.e. with no visible enhancing residual tumor. The data originates from 12 different hospitals in Europe and the US, and the data origin and distribution across hospitals has been described in a previous study [9]. Study protocols were approved by local regional ethics committees in the respective countries of data origin.

Test set All 73 patients from one hospital was kept outside training as an external test set, to ensure an unbiased evaluation of the generalizability of the models. The external test set included a subset of 20 patients annotated by 8 annotators with different levels of experience, grouped into 4 novice and 4 expert annotators. This subset, hereby referred to as the inter-rater test set, was used in a previous study for evaluating the inter-rater agreement of manual post-operative segmentation [19], and was used to compare the models performance against the performance of human annotators. Additionally, all 73 patients in the entire external test set had been previously annotated by an independent expert, which were used as ground truth for the main evaluation and comparison of experiments, as the size of the inter-rater test set was quite limited.

2.2 Methods

All experiments were conducted using a 3D patch-wise (PW) AGU-Net architecture [3]. This architecture was selected as a basis model because it was shown in a previous study [9] that this architecture achieved a similar segmentation performance as human expert raters and the nnU-Net, while also achieving a reasonable classification performance. The Attention component of the AGU-Net was also deemed important for the network to locate the area of the tumor. An overview of the proposed method is shown in Figure 1. The patients EPMR T1w-CE and T1w sequences were used as input for training as adding more sequences was shown to not improve segmentation performances [9].

Sampling strategies for handling the extreme class imbalance One of the main challenges of early post-operative segmentation is the extreme class imbalance between the residual tumor tissue class and the background. Indeed, 35% of the patients in the dataset were defined as GTR, having no enhancing residual tumor. Among the remaining 65% of the patients, most tumors were extremely small with an average tumor volume of 3 ml. For comparison, the average pre-operative tumor volume lies around 35 ml. The class imbalance increases when training a patch-wise model as more patches not overlapping with the residual tumor will lead to an even higher portion of data samples with no positive voxels. In experiment 1-5, different sampling strategies for handling the class imbalance were investigated.

Impact of network levels and kernel size Another main challenge of early post-operative segmentation is the thin and fragmented structure of the residual tumors, generally spread around the resection cavity. In addition, the resolution of the input images is incrementally reduced by a factor of two for each level down the encoder path of U-Net shaped architectures. As such, small structures with a thickness of only a few voxels will rapidly disappear without contributing to the deeper architecture's feature maps. To better capture the small structures, a network with larger kernel sizes and fewer encoding levels might be of interest (cf. experiments 6 and 7).

Comparison with human inter-rater variability The Dice metric has well-known limitations when segmenting small or non-existent structures [16], and few reference works or benchmarks for early post-operative segmentation have been published. Therefore, a thorough assessment of realistic expectations for model performance, compared against human annotators, was deemed imperative. To this end, the annotations from the inter-rater test set were used to generate different consensus agreement annotations. An average annotation of human raters presents the benefit to minimize the inter-rater variability when used to benchmark a models segmentation and classification performance, while at the same time assessing the inter-rater agreement amongst all annotators. A consensus agreement was created for each of the two groups of raters (i.e., experts and novices), and one for the ensemble of all raters. A voxel was counted as positive if it had been annotated by more than half of the annotators, e.g. by at least 3 on a group level, and at least 5 for all annotators. In addition, a union of all annotations was created, where all voxels annotated by at least one annotator were counted as positive. Dice scores were computed for the consensus agreement and union annotations using the independent single rater ground truth annotations as reference. The purpose of this was two-fold: to contrast the model performance against human rater performance using a reference completely independent of both, and to assess the quality of the single rater ground truth annotations. Finally, the Dice score was computed for the different consensus annotations using the union of annotations as a reference, in order to assess the agreement of the majority votes and the union of all raters.

Experiments A total of seven experiments were conducted in this work, as summarized in the following.

1. PW_AGU_pos: Train on positive (with residual tumor) patches of size 160^3 voxels.
2. PW_AGU_all: Train on all available samples (i.e., positive and negative).
3. PW_AGU_fine Fine-tuning *PW_AGU_post*-model on all available samples.
4. PW_AGU_rand: Patch-size set to 96^3 voxels and training with sampling only one patch randomly from each patient during each epoch.
5. PW_AGU_tumor: Sample 50% of the patches centered around a tumor lesion, and 50% at random. One patch was selected from each patient in each epoch.
6. PW_AGU_ker7: AGU-Net model with two levels and a 7-voxel kernel size.
7. PW_AGU_ker5: AGU-Net model with three levels and a 5-voxel kernel size.

3 Results

3.1 Implementation details

The study was implemented in Python 3.8 using Tensorflow 2.8.0, on a machine with a Tesla V100S (32GB) GPU. All models were trained from scratch for 300 epochs with early stopping (patience 15), batch size of 2 to 4, and accumulated gradients [15] of 8 to 16, always giving an effective batch size of 32. The Dice loss was used with the Adam optimizer, using an initial learning rate of 10^{-4}. As data augmentation, random flipping, rotation, and translation were applied. The implementation and trained weights of the best trained model are available through the open source software Raidionics [4], and the code for validation is available on Github [2]. The data used in this project is not openly available due to patient privacy, but access can be granted through collaborative projects.

Table 1. Results for all experiments, compared against patch-wise (PW) nnU-Net, and full volume (FV) AGU-Net, reported over an external test set of 73 patients.

Model	Segmentation		Classification		
	Dice	HD95 (mm)	Sens.	Spec.	bAcc
PW_AGU_post	49.78±25.87	22.93±34.26	86.27	36.36	61.32
PW_AGU_all	46.58±27.88	23.61±39.18	82.35	**50.00**	66.18
PW_AGU_fine	48.18±27.28	23.61±35.13	86.27	45.45	65.86
PW_AGU_rand	50.10±25.20	22.52±34.06	88.24	**50.00**	69.12
PW_AGU_tumor	49.95±26.48	**20.78±33.99**	90.20	**50.00**	**70.10**
PW_AGU_ker7	47.14±25.39	27.67±37.48	88.24	36.36	62.30
PW_AGU_ker5	**51.09±23.76**	25.29±33.72	**92.16**	27.27	59.71
PW_nnU-Net	60.08±21.09	20.18±32.15	96.08	18.18	57.13
FV_AGU-Net	45.04±28.21	20.65±35.57	82.35	63.64	72.99

3.2 Experiment results

Experimental results on sampling strategies, architecture levels, and kernel sizes are summarized in Table 1. The models' segmentation performances was evaluated using the Dice score and 95% Hausdorff distance (HD95). Additionally, the models were evaluated on the auxiliary task of classification of residual tumor (RT) and gross total resection (GTR), based on thresholding of the predicted tumor volumes. The classification performance was evaluated in terms of the Balanced Accuracy (bAcc), which is the mean of the Sensitivity (Sens.) and Specificity (Spec.), to account for the class imbalance. Results obtained with previously trained models [9] (i.e., patch-wise nnU-Net and full-volume AGU-Net) have been included for reference. Dice scores are reported only for the patients with residual tumor according to the ground truth annotation. Most experiments resulted in similar average Dice scores and Hausdorff distances, but the classification results varied more across the experiments. None of the patch-wise AGU-Net models achieved similar segmentation performance to the nnU-Net. However, the high rate of false positives and low patient-wise specificity make nnU-Net models unsuitable for use in a clinical setting. All models achieved slightly higher Dice scores than the full-volume AGU-Net model, although the classification performances were lower due to lower specificity. Experiment 5 delivered the top-performing model, achieving a high Dice score and low Hausdorff distance. Overall, model 5 represents the best trade-off between patient-wise sensitivity and specificity, with the highest balanced accuracy (bAcc) of 70%. To summarize, the sampling strategies for handling the extreme class imbalance only led to incremental improvements in the Dice scores. Nevertheless, the trade-off between sensitivity and specificity for classification seemed to improve slightly with smaller patches and higher sampling frequency of the area encompassing the residual tumor (*PW_AGU_tumor*). The last experiment with fewer encoding levels and kernel size 5 achieved the highest Dice score and patient-wise sensitivity out of all experiments at the cost of a low specificity. Architecture and hyper-parameters optimization appear to provide only marginal performance improvement. While ensembling schemes leveraging predictions from multiple models has shown an increase in Dice score, the cost often is a worse trade-off between sensitivity and specificity. The reported nnU-Net results in the table were obtained from ensembling five models. The nnU-Net framework, the cornerstone of all top-performing submissions to the BraTS challenge over the last 4 years, optimizes the hyperparameters of the U-Net architecture and pre-processing, for the particular dataset at hand. As shown by the 2023 winning submission, having access to a larger training dataset through synthetic data generation, in addition to ensembling of an large number of models, yielded the best performance. However, both ensembling and generation of synthethic data is costly both in terms of computational resources and runtime, making trained models less affordable for clinics. The quality of the annotations and the definition of the gold standard is clearly a limiting factor for performance in the early post-operative case, which will be analysed more in-depth in the comparison with the human inter-rater variability.

Table 2. Results for PW_AGU_tumor on the inter-rater test set, compared against PW_nnU-Net, FV_AGU-Net, and the consensus agreement annotations.

Experiment	Segmentation		Classification		
	Dice	HD95 (mm)	Sens.	Spec.	bAcc
PW_AGU_tumor	48.92±28.30	13.52±9.88	90.00	60.00	75.00
PW_nnU-Net	55.37±23.14	17.06±15.07	90.00	30.00	60.00
FV_AGU-Net	49.78±28.36	14.56±10.67	80.00	80.00	80.00
consensus-novices	30.48±29.57	18.56±12.52	60.00	90.00	75.00
consensus-experts	50.03±27.11	9.15±5.88	80.00	80.00	80.00
consensus-all-annotators	44.86±30.48	10.09±6.80	70.00	90.00	80.00

3.3 Comparison with human inter-rater variability

The results over the inter-rater test set for *PW_AGU_tumor*, *PW_nnU-Net*, *FV_AGU-Net*, and the consensus agreement annotations are reported in Table 2. The model from *PW_AGU_tumor* outperforms the consensus agreement annotation for the novices and the ensemble of all raters in terms of Dice, and performs at a similar level as the expert consensus annotations. The human raters all have a higher patient-wise specificity than the model from *PW_AGU_tumor*, but the model has a higher sensitivity. The discrepancy between the human raters in terms of Dice score illustrates the level of difficulty of the task. While the Dice score is fundamental for assessing a segmentation model's performance, clear limitations exist for evaluating a model's usefulness in clinical practice. Therefore, models' performance should also take into account more clinically relevant metrics such as volume agreement or the ability to classify patients with residual tumor and gross total resections.

In Figure 2, the predicted volumes from *PW_AGU_tumor* are plotted against 3 different reference volumes: the ground truth annotations, the consensus agreement annotations, and the union of all raters annotations. The correlation is high between the model and both of the first two references, but the agreement seems to be higher with the consensus agreement annotation as the slope is closer to 1. Using the consensus agreement annotation as a reference, the average Dice score with the model predictions was 0.49 vs 0.28 for the union of all annotations, indicating a high agreement between the model and consensus agreement compared to the inter-rater agreement. Annotation by consensus agreement of multiple experts are usually considered to be of higher quality than single annotations. Models showing better agreement with the consensus agreement annotations should therefore be considered to achieve higher performance, illustrating robustness and ability to generalize. Ground truth annotations produced by single raters are one of the major limitations of any training data. Models should ideally be trained on consensus agreement annotations from multiple experts, but unfeasible in practice from the time and effort required for annotating 3D MRI scans. Example predictions from *PW_AGU_tumor*, annotations and Dice scores are shown in Figure 3, illustrating the challenges with small and fragmented lesions as well as the discrepancy in the human raters annotations.

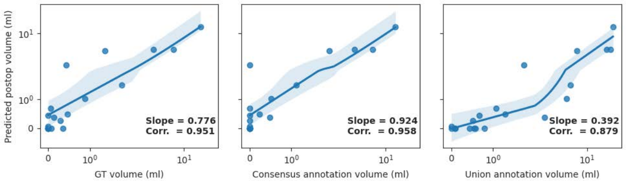

Fig. 2. Linear regression plots of predicted volume vs (1) GT (2) consensus agreement, and (3) union of annotations. The plots are shown with log-axis for visibility.

4 Conclusions

In this study, experiments on sampling strategies and architecture designs to increase segmentation and classification performance for early post-operative glioblastoma segmentation were presented. Improving segmentation performance

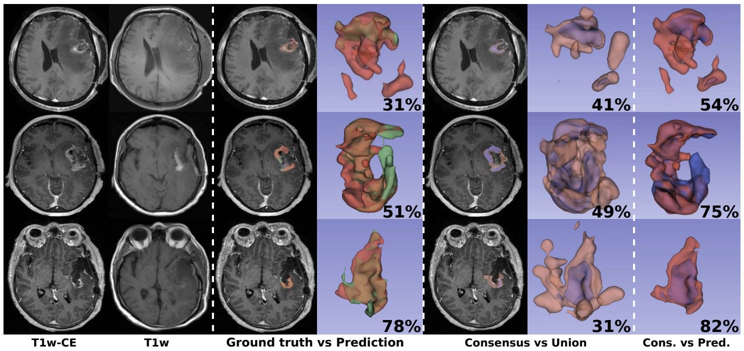

Fig. 3. Examples with Dice scores (%). Ground truth annotations in green, predictions in red, consensus annotation in blue, and union annotations in pink.

and patient-wise sensitivity was shown to be challenging without compromising on the specificity and classification performance. Part of the explanation can be linked to the quality and robustness of the manual annotations. The level of difficulty for the task requires experienced raters, but inter-rater variability was highlighted as an ensemble of clinical expert raters disagrees on the segmentation of the residual tumor. In spite of these challenges, our model shows robustness and high agreement with the ensemble of clinical expert raters. The automatic method for early post-operative segmentation could therefore be a time-saving and accurate alternative to the current standard manual method for measuring residual tumor volume after surgery of glioblastoma.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Berntsen, E.M., et al.: Volumetric segmentation of glioblastoma progression compared to bidimensional products and clinical radiological reports. Acta Neurochirurgica **162**(2), 379–387 (2020). https://doi.org/10.1007/s00701-019-04110-0
2. Bouget, D.: dbouget/validation_metrics_computation: v1.0.0 (2024), https://github.com/dbouget/validation_metrics_computation
3. Bouget, D., Pedersen, A., Hosainey, S.A.M., Solheim, O., Reinertsen, I.: Meningioma Segmentation in T1-Weighted MRI Leveraging Global Context and Attention Mechanisms. Frontiers in Radiology **1**(September) (2021). https://doi.org/10.3389/fradi.2021.711514
4. Bouget, D., et al.: Preoperative brain tumor imaging: Models and software for segmentation and standardized reporting. Frontiers in Neurology **13** (2022). https://doi.org/10.3389/fneur.2022.932219, https://www.frontiersin.org/articles/10.3389/fneur.2022.932219
5. Coburger, J., et al.: Impact of extent of resection and recurrent surgery on clinical outcome and overall survival in a consecutive series of 170 patients for glioblastoma in intraoperative high field magnetic resonance imaging. Journal of neurosurgical sciences **61**(3), 233–244 (6 2017). https://doi.org/10.23736/S0390-5616.16.03284-7

6. Davis, M.E.: Glioblastoma: Overview of disease and treatment. Clinical Journal of Oncology Nursing **20**(5), 1–8 (10 2016). https://doi.org/10.1188/16.CJON.S1.2-8
7. Ferreira, A., et al.: How we won BraTS 2023 Adult Glioma challenge? Just faking it! Enhanced Synthetic Data Augmentation and Model Ensemble for brain tumour segmentation pp. 1–18 (2024), http://arxiv.org/abs/2402.17317
8. Ghaffari, M., et al.: Automated post-operative brain tumour segmentation: A deep learning model based on transfer learning from pre-operative images. Magnetic Resonance Imaging **86**(August 2021), 28–36 (2022). https://doi.org/10.1016/j.mri.2021.10.012, https://doi.org/10.1016/j.mri.2021.10.012
9. Helland, R.H., Ferles, A., et al.: Segmentation of glioblastomas in early post-operative multi-modal MRI with deep neural networks. Scientific Reports **13**(1), 1–13 (2023). https://doi.org/10.1038/s41598-023-45456-x, https://doi.org/10.1038/s41598-023-45456-x
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z, http://dx.doi.org/10.1038/s41592-020-01008-z
11. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-Net for Brain Tumor Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12659 LNCS**, 118–132 (2021). https://doi.org/10.1007/978-3-030-72087-2–“˙”11
12. Kundu, S., et al.: Ase-net for segmentation of post-operative glioblastoma and patient-specific fine-tuning for segmentation refinement of follow-up mri scans. SN Computer Science **5**(1), 106 (2023)
13. Lotan, E., et al.: Development and Practical Implementation of a Deep Learning-Based Pipeline for Automated Pre- and Postoperative Glioma Segmentation. American Journal of Neuroradiology **43**(1), 24–32 (2022). https://doi.org/10.3174/ajnr.A7363
14. Menze, B.H., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
15. Pedersen, A., Bouget, D.: andreped/GradientAccumulator: v0.3.1 (2023). https://doi.org/10.5281/zenodo.7582309, https://doi.org/10.5281/zenodo.7582309
16. Reinke, A., Tizabi, M.D., Sudre, C.H., Eisenmann, M., et al., R.: Common Limitations of Image Processing Metrics: A Picture Story (2021), http://arxiv.org/abs/2104.05642
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (2015). https://doi.org/10.1007/978-3-319-24574-4–“˙”28
18. Skaga, E., et al.: Real-world validity of randomized controlled phase III trials in newly diagnosed glioblastoma: To whom do the results of the trials apply? Neuro-Oncology Advances **3**(1), 1–12 (2021). https://doi.org/10.1093/noajnl/vdab008
19. Visser, M., et al.: Inter-rater agreement in glioma segmentations on longitudinal MRI. NeuroImage: Clinical **22**(July 2018), 101727 (2019). https://doi.org/10.1016/j.nicl.2019.101727, https://doi.org/10.1016/j.nicl.2019.101727
20. Wen, P., et al.: Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group (4 2010). https://doi.org/10.1200/JCO.2009.26.3541

21. Zeineldin, R.A., Karar, M.E., Burgert, O., Mathis-Ullrich, F.: Multimodal CNN Networks for Brain Tumor Segmentation in MRI: A BraTS 2022 Challenge Solution pp. 1–13 (2022), http://arxiv.org/abs/2212.09310