scientific reports

Check for updates

OPEN

Segmentation of glioblastomas in early post-operative multi-modal MRI with deep neural networks

Ragnhild Holden Helland[1,2,28]✉, Alexandros Ferles[3,4,28], André Pedersen[1], Ivar Kommers[3,5], Hilko Ardon[6], Frederik Barkhof[4,7], Lorenzo Bello[8], Mitchel S. Berger[9], Tora Dunås[10], Marco Conti Nibali[11], Julia Furtner[12,13], Shawn Hervey-Jumper[9], Albert J. S. Idema[14], Barbara Kiesel[15], Rishi Nandoe Tewari[16], Emmanuel Mandonnet[17], Domenique M. J. Müller[3,5], Pierre A. Robe[18], Marco Rossi[19], Lisa M. Sagberg[20,21], Tommaso Sciortino[11], Tom Aalders[22], Michiel Wagemakers[23], Georg Widhalm[15], Marnix G. Witte[24], Aeilko H. Zwinderman[25], Paulina L. Majewska[18,26], Asgeir S. Jakola[10,27], Ole Solheim[18,26], Philip C. De Witt Hamer[3,5], Ingerid Reinertsen[1,2], Roelant S. Eijgelaar[3,5,29] & David Bouget[1,29]

Extent of resection after surgery is one of the main prognostic factors for patients diagnosed with glioblastoma. To achieve this, accurate segmentation and classification of residual tumor from post-operative MR images is essential. The current standard method for estimating it is subject to high inter- and intra-rater variability, and an automated method for segmentation of residual tumor in early post-operative MRI could lead to a more accurate estimation of extent of resection. In this study, two state-of-the-art neural network architectures for pre-operative segmentation were trained for

[1]Department of Health Research, SINTEF Digital, 7465 Trondheim, Norway. [2]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway. [3]Cancer Center Amsterdam, Brain Tumor Center, Amsterdam University Medical Centers, 1081 HV Amsterdam, The Netherlands. [4]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Vrije Universiteit, 1081 HV Amsterdam, The Netherlands. [5]Department of Neurosurgery, Amsterdam University Medical Centers, Vrije Universiteit, 1081 HV Amsterdam, The Netherlands. [6]Department of Neurosurgery, Twee Steden Hospital, 5042 AD Tilburg, The Netherlands. [7]Institutes of Neurology and Healthcare Engineering, University College London, London WC1E 6BT, UK. [8]Neurosurgical Oncology Unit, Department of Oncology and Hemato-oncology, Humanitas Research Hospital, Università Degli Studi di Milano, 20122 Milan, Italy. [9]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA 94143, USA. [10]Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, 405 30 Gothenburg, Sweden. [11]IRCCS Ospedale Galeazzi Sant'Ambrogio, 20157 Milan, Italy. [12]Department of Biomedical Imaging and Image-guided Therapy, Medical University Vienna, 1090 Vienna, Austria. [13]Research Center for Medical Image Analysis and Artificial Intelligence (MIAAI), Faculty of Medicine and Dentistry, Danube Private University, 3500 Krems, Austria. [14]Department of Neurosurgery, Northwest Clinics, 1815 JD Alkmaar, The Netherlands. [15]Department of Neurosurgery, Medical University Vienna, 1090 Vienna, Austria. [16]Department of Neurosurgery, Haaglanden Medical Center, 2512 VA The Hague, The Netherlands. [17]Department of Neurological Surgery, Hôpital Lariboisière, 75010 Paris, France. [18]Department of Neurology and Neurosurgery, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands. [19]Department of Medical Biotechnology and Translational Medicine, Università Degli Studi di Milano, 20122 Milan, Italy. [20]Department of Neurosurgery, St. Olavs hospital, Trondheim University Hospital, 7030 Trondheim, Norway. [21]Department of Public Health and Nursing, Norwegian University of Science and Technology, 7491 Trondheim, Norway. [22]Department of Neurosurgery, Isala, 8025 AB Zwolle, The Netherlands. [23]Department of Neurosurgery, University Medical Center Groningen, University of Groningen, 9713 GZ Groningen, The Netherlands. [24]Department of Radiation Oncology, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands. [25]Department of Clinical Epidemiology and Biostatistics, Amsterdam University Medical Centers, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands. [26]Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway. [27]Department of Neurosurgery, Sahlgrenska University Hospital, Gothenburg, Sweden. [28]These authors contributed equally: Ragnhild Holden Helland and Alexandros Ferles. [29]These authors jointly supervised this work: Roelant S. Eijgelaar and David Bouget. ✉email: ragnhild.h.helland@ntnu.no

the task. The models were extensively validated on a multicenter dataset with nearly 1000 patients, from 12 hospitals in Europe and the United States. The best performance achieved was a 61% Dice score, and the best classification performance was about 80% balanced accuracy, with a demonstrated ability to generalize across hospitals. In addition, the segmentation performance of the best models was on par with human expert raters. The predicted segmentations can be used to accurately classify the patients into those with residual tumor, and those with gross total resection.

Glioblastoma, the most common malignant primary brain cancer, requires a multidisciplinary treatment approach comprising maximum safe surgical resection, followed by concurrent radiation and chemotherapy[1]. Even so, median survival in unselected patients is only 12 months[2]. Due to the high invasiveness, a complete resection of all tumor cells is not possible. Still, extensive surgical resections are associated with longer survival[3], but as surgically induced neurological impairment is associated with shorter survival[4], extent of resection (EOR) and surgical strategy, for example resection or biopsy only, needs to be weighed up against risks in individual patients.

The EOR is calculated as the ratio between the surgically-removed and pre-operative tumor volume, which relies on an accurate segmentation of the tumor tissue in both pre- and post-operative MR scans. In recent years, a large body of work has focused exclusively on automated segmentation of pre-operative glioblastoma, yet the task of residual tumor segmentation from early post-operative MRI (EPMR) has gained less attention from the research community. In current practice, the residual tumor size is estimated manually through eye-balling[5], or using crude measures such as the bi-dimensional product of the largest axial diameter of the contrast enhancing residual tumor, according to the Response Assessment in Neuro-Oncology (RANO) criteria[6]. Manual volume segmentations are more sensitive but expertise-dependent and time-consuming, with high inter- and intra-rater variability[5,7]. An automated method for post-operative tumor volume segmentation from EPMR would therefore be beneficial.

Glioblastoma segmentation from pre-operative MR scans has received a lot of attention in the literature in recent years. Many contributions were motivated by the MICCAI Brain Tumor Segmentation (BraTS) Challenge[8]. With the emergence of fully convolutional neural networks (CNNs)[9], deep learning-based approaches have nearly completely replaced more conventional methods in medical image segmentation[10]. Variants of the U-Net architecture[11] have facilitated the basis-architecture for the majority of auto-segmentation algorithms, including DeepMedic[12], Attention U-Net[13], and the recently established nnU-Net[14], with state-of-the-art results in several medical image segmentation benchmarks. The winning submissions in the BraTS challenge in 2021 and 2022 were an extension of the nnU-Net architecture[15], and an ensemble of three state-of-the art architectures for medical image segmentation, comprising nnU-Net, DeepSeg, and DeepSCAN[16], respectively. In the absence of a publicly available dataset for residual tumor segmentation from EPMR, the literature on this problem is sparse when compared to the pre-operative segmentation task. Semi-automatic methods, combining of one or several voxel- or shape-based image segmentation algorithm, have been proposed from intensity thresholding (e.g., Otsu and relative entropy)[17–19], fuzzy algorithms[18], Gaussian mixture model[20], morphological operations[19], region-based active contours[21,22], the level set approach[21–23], and CNNs[24]. Unfortunately, these methods relied on user inputs, either by manual initialisation, or by interactive refinement of the resulting segmentation. They are therefore challenging to use in clinical practice, and in large datasets. In addition, all validation studies were solely performed on single-center local datasets, consisting of 15 to 37 patients, making if difficult to demonstrate the generalizability of the proposed methods.

Regarding fully automated approaches, Meier et al.[25] presented an automated method based on decision forests for residual tumor segmentation using EPMR from 19 patients. A more recent work by Ghaffari et al.[26] proposed to fine-tune a 3D densely connected U-Net, pre-trained on the BraTS20 dataset, on a local dataset of 15 post-operative glioblastomas. However, the MR scans were all acquired for radiation therapy planning and not within the recommended time frame to acquire EPMR scans, within 72 hours after surgical resection[6]. Deep learning approaches have recently shown to outperform more traditional algorithms on most image segmentation tasks, including segmentation of pre-operative glioblastomas[15,16]. The utmost requirement is the number of included patients and the quality of the MR images comprising a study dataset. Preferably, the data should originate from different locations, to evaluate the ability of the trained models to generalize across different hospitals, scanners, or clinical practice.

In this work, we determine the performance of two CNN architectures to segment residual enhancing glioblastoma on early post-operative scans. The selected architectures are the nnU-Net, state-of-the-art for pre-operative glioblastoma segmentation, and AGU-Net, an architecture developed for pre-operative segmentation of brain tumors. These architectures have both demonstrated excellent performance on pre-operative segmentation in previous studies on pre-operative brain tumor segmentation[27–29], and they exhibit different strengths and weaknesses. The automatic results are compared with manual segmentations, using different combinations of MRI scans in a large dataset consisting of paired pre- and early post-operative MRI scans from 956 patients in 12 medical centers in Europe and the United States. Extensive validation studies are presented to identify the best architecture configuration, quantify the performances and ability to generalize, and highlight potential relevance for use in clinical practice. Finally, the best performing models are made publicly available and integrated into the open software Raidionics[29].

Materials & Method

Ethics and informed consent statement

The study was conducted in accordance with the Declaration of Helsinki. The study protocol was approved by the Medical Ethics Review Committee of VU University Medical Center (IRB00002991, 2014.336), and the Norwegian regional ethics committee (REK ref. 2013/1348 and 2019/510). Written informed consent was obtained from patients as required for each participating hospital.

Data

A dataset comprised of pre-operative and early post-operative MRI scans from 956 patients, who underwent surgical resection of glioblastoma, was assembled for this study. Twelve different hospitals across Europe and in the US contributed data, with the following patient distribution per center: 23 patients from the Northwest Clinics, Alkmaar, Netherlands (*ALK*); 73 patients from the Amsterdam University Medical Centers, location VU medical center, Netherlands (*AMS*); 43 patients from the University Medical Center Groningen, Netherlands (*GRO*); 40 patients from the Medical Center Haaglanden, the Hague, Netherlands (*HAG*); 55 patients from the Humanitas Research Hospital, Milano, Italy (*MIL*); 41 patients from the Hôpital Lariboisière, Paris, France (*PAR*); 108 patients from the University of California San Francisco Medical Center, U.S. (*SFR*); 53 patients from the University Medical Center Utrecht, Netherlands (*UTR*); 45 patients from the Medical University Vienna, Austria (*VIE*); 51 patients from the Isala hospital, Zwolle, Netherlands (*ZWO*); 237 patients from St. Olavs hospital, Trondheim University Hospital, Norway (*STO*); and 187 patients from the Sahlgrenska University Hospital, Gothenburg, Sweden (*GOT*).

The cohorts are subsets of a broader dataset, thoroughly described previously for their pre-operative content[30], for patients with available EPMR data. All EPMR scans were acquired within 72 hours after surgery, with the exception of the UTR center where the limit used was up to one week post-surgery. The recommended time frame for acquiring the EPMR scans has been stated in the National Comprehensive Cancer Network (NCCN) recommendations[31], in order to maximize differences between residual enhancing tumor and enhancement due to post-surgical changes in the tissue[32,33]. For each patient in the dataset, the following post-operative MRI sequences were acquired: T1-weighted (T1w), gadolinium-enhanced T1-weighted (T1w-CE), and T2-weighted fluid attenuated inversion recovery (FLAIR).

For the T1w-CE sequence, the volume dimensions are covering $[128;896] \times [42;896] \times [17;512]$ voxels, and the voxel size ranges are $[0.26;1.2] \times [0.26;5.0] \times [0.49;7.2]$ mm^3. An average T1w-CE volume has a resolution of [430, 461, 180] voxels with a spacing of $[0.67 \times 0.64 \times 1.96]$ mm^3. Details about the the resolution and spacing of the other sequences can be found in the supplementary materials, Table S2.

The residual tumor tissue was manually segmented in 3D in T1w-CE MR scans by trained annotators, supervised by expert neuroradiologists and neurosurgeons. The manual segmentation was performed using all available standard MR sequences, and residual tumor tissue was defined as enhancing tissue in the T1w-CE scan, but darker in the T1w scan. Hence, blood was distinguished from residual tumor by a hyperintense signal on T1w scans. For each patient, a further post-operative distinction can be made between cases showcasing residual tumor (RT) in EPMR scans and cases presenting a gross total resection (GTR), defined as a residual tumor volume of less than 0.175 ml[34]. The cut-off was chosen to reduce risk of interpretation problems when distinguishing between tumour enhancement and that of non-specific enhancement, such as small vessels or enhancing pia mater. Under this paradigm, 352 patients (35%) in our dataset had a GTR, whereas the remaining 604 patients had residual tumor. The average post-operative tumor volume is 3 ml, whereas the average pre-operative tumor volume is 35 ml. An overview of the data from the 12 hospitals is shown in Table 1, and some examples are illustrated in Fig. 1.

In addition, 20 patients out of the 73 in the AMS cohort have been annotated eight times in total, by four novices raters and four experts raters. This cohort has been used in a previous study to evaluate the inter-rater

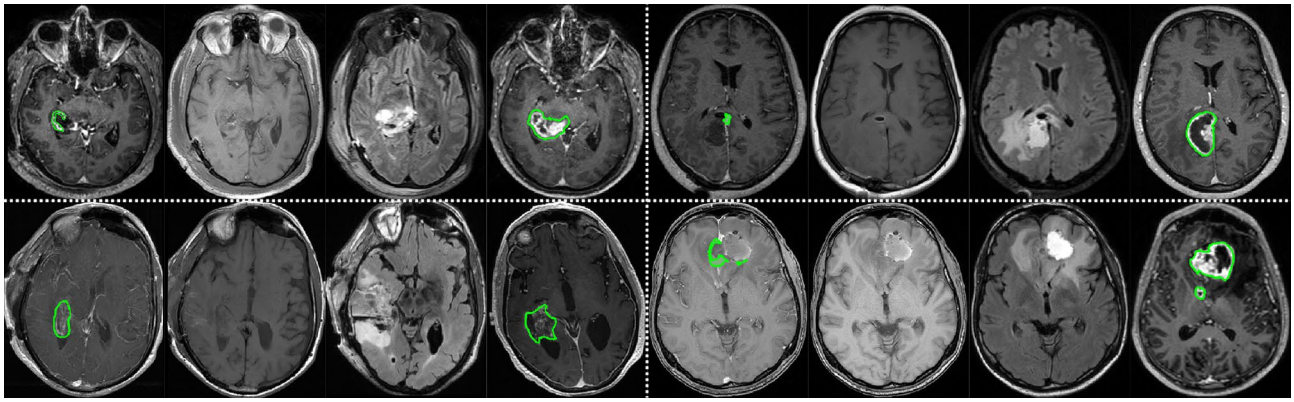

Figure 1. Dataset examples for four patients, separated by white dash-lines. For each patient, an axial view from the EPMR T1w-CE, EPMR T1w, EPMR FLAIR, and pre-operative T1w-CE are displayed. Outlines of the manually annotated tumors are shown in green.

Hospital	HAG	MIL	ZWO	VIE	ALK	PAR	SFR	GRO	UTR	AMS	STO	GOT
Patients	40	55	51	45	23	41	108	43	53	73	237	187
RT	23	34	18	30	18	29	80	26	20	51	162	113
GTR	17	21	33	15	5	12	28	17	33	22	75	74
RT ratio (%)	57.5	61.8	35.3	66.7	78.3	70.7	74.1	60.5	37.7	69.9	68.4	60.4

Table 1. Dataset distributions and statistics across the twelve hospitals, represented by their acronyms. RT: residual tumor, GTR: gross total resection.

variability of tumor annotation from annotators with different levels of experience[7], and will be referred to as the inter-rater variability dataset in the remainder of the document.

Segmentation process

Similar to our previous work on pre-operative glioblastoma segmentation[27], the following two competitive CNN architectures were selected for the task of voxel-wise segmentation of residual tumor tissue: patch-wise nnU-Net[14] and full-volume AGU-Net[28].

Multiple MR sequences combinations can be considered as input for the CNN architectures. In an attempt to minimize essential input and following typical incremental assessment by neuroradiologists, these combinations of input sequences were considered for the automated segmentations of post-operative tumor: (**A**) the EPMR T1w-CE scan only, (**B**) the EPMR T1w-CE and EPMR T1w, to potentially distinguish between blood and residual tumor, (**C**) all standard EPMR sequences: T1w-CE, T1w, and FLAIR scans, (**D**) the EPMR T1w-CE and EPMR T1w, and the pre-operative T1w-CE MR scan and corresponding tumor segmentation mask, and (**E**) all standard EPMR sequences: T1w-CE, T1w, and FLAIR scans, and the pre-operative T1w-CE MR scan and corresponding tumor segmentation mask. An overview of the whole segmentation pipeline with the different input designs and subsequent steps is presented in the following sections, and illustrated in Fig. 2.

Pre-processing

For proper anatomical consistency across the different inputs sequences, an initial image-to-image registration procedure was performed. The EPMR T1w-CE scan was elected as the reference space and all subsequent volumes were registered to it using the SyN diffeomorphic method[35] from the Advanced Normalization Tools (ANTs) framework[36]. Skull-stripping was subsequently performed on all input MR scans, based on the brain mask from the EPMR T1w-CE scan. All brain masks were automatically generated using a pre-trained slab-wise AGU-Net model with input shape $256 \times 192 \times 32$ voxels. For the nnU-Net architecture, the pre-processing was automatically decided by the framework based on the dataset, and all inputs were resampled to $0.5 \times 0.5 \times 1.0\,mm^3$ spacing and zero-mean normalized. For the AGU-Net architecture, the full-resolution analysis required a lower resolution, and therefore all inputs were resampled to an isotropic $1.0\,mm^3$ spacing, resized to $128 \times 128 \times 144$ voxels, and zero-mean normalized.

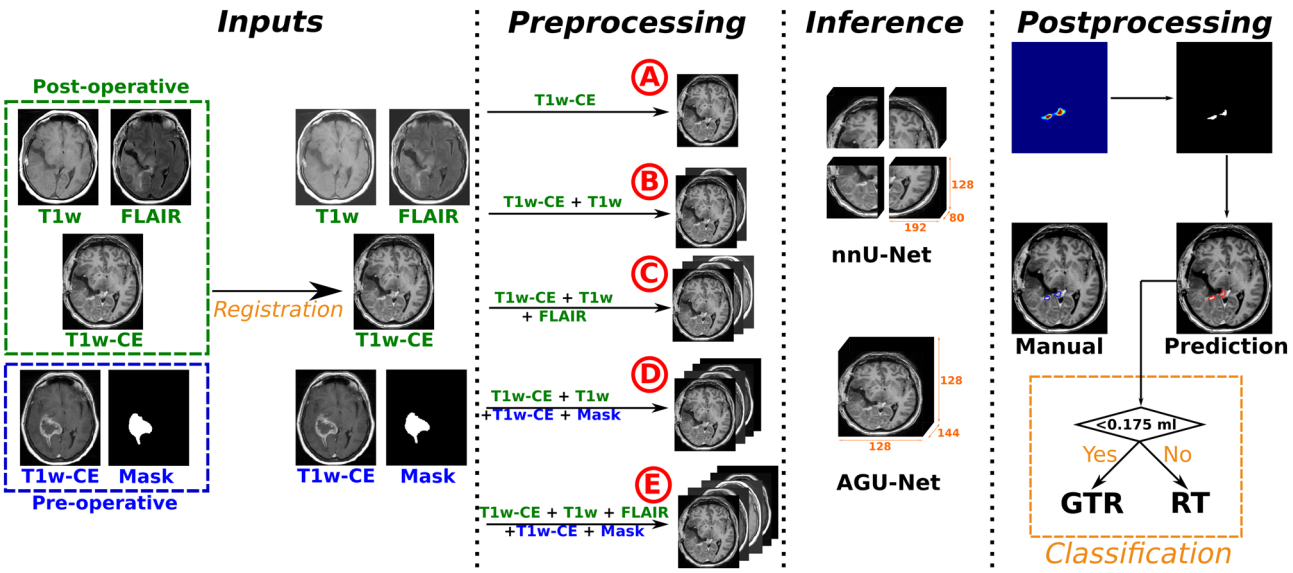

Figure 2. Overall residual tumor segmentation pipeline from EPMR scans and classification between gross total resection or residual tumor. The registration is performed using the SyN approach from ANTs, multiple input configurations using different combinations of MR sequences were considered (noted from A to E), and two architectures were evaluated: nnU-Net and AGU-Net.

Training specifications for the nnU-Net architecture

Architecture design. From the nnU-Net framework analysis of the dataset, the 3D full-resolution U-Net with the following parameters was recommended, using $192 \times 128 \times 80$ voxels as input patch size. The network used five levels, downsampling using strided convolution layers, and upsampling using transposed convolution layers. Kernel size of $1 \times 3 \times 3$ voxels for the first level, $3 \times 3 \times 3$ for the remaining four levels, and filter sizes of $\{32, 64, 128, 256, 320\}$ were used for each level, respectively. The loss function was a combination of the Dice score and cross-entropy. A stride of one was used for the convolution layers.

Network training. All models were trained from scratch for 1000 epochs using a stochastic gradient descent with Nesterov momentum optimizer (momentum=0.99). One epoch was defined as 250 batch iterations with a batch size of two. On-the-fly data augmentations were performed comprising rotation, scaling, additive Gaussian noise, Gaussian blur, brightness and contrast augmentation, and gamma augmentation.

Training specifications for the AGU-Net architecture

Architecture design. The AGU-Net, as described by Bouget et al.[28], is a 3D U-Net architecture with an integrated attention-gated mechanism, with five block levels using filter sizes of {16, 32, 128, 256, 256}, respectively. The input size of the network was set to $128 \times 128 \times 144 \times S$, with S being the number of sequences used as input. The architecture also uses multi-scale input and deep supervision. The class-averaged Dice loss, excluding the background, was used for training the different models.

Network training. All models were initialized using pre-trained weights from the best pre-operative glioblastoma segmentation model[27], and only the input layer was adapted to account for the different input combinations considered. The Adam optimizer was used with an initial learning rate of 1×10^{-3}, and the training was stopped after 30 consecutive epochs without validation loss improvement. Gradient accumulation[37] was performed to increase the batch size from 2 samples to 32, tackling graphics processing unit (GPU) memory limitations for large batch training. Data augmentation techniques were leveraged including horizontal and vertical flipping, random rotations in the range $[-20°, 20°]$, and a translation of up to 10% of the axis dimension. Each augmentation was performed with a probability of 50% for each training sample.

Post-processing and GTR classification

During inference, residual tumor tissue was predicted by each trained model, resulting in a probability map of the same resolution as the EPMR T1w-CE scan. A binary mask was then generated from the probability map, using the best threshold determined from the validation studies. The binary mask was further refined by filtering out potential noise, inherent to the voxel-wise segmentation task, by applying a connected components analysis and removing any identified object smaller than 20 voxels. Finally, the refined binary mask was used to assess whether gross total resection has been achieved for the patient.

Validation studies

In this work, the trained models were assessed based on their ability to perform segmentation of the residual tumor and to classify patients into those with gross total resection and those with residual tumor. For the segmentation task, only two classes are considered, whereby a positive voxel exhibits tumor tissue, whereas a negative voxel represents either background or normal tissue. For the classification task, a rest tumor volume threshold was selected to serve as cut-off value.

Protocols

The validation studies presented in this work were conducted following a five-fold cross-validation, summarized in Table 2. First, all patients from 11 out of the 12 the hospitals in our dataset, excluding the AMS cohort, were split into five hospital-stratified folds, with an approximately balanced number of patients in each fold. The remaining 73 patients from the AMS hospital were kept as an hold-out test set. For each iteration of the cross-validation, four folds were used for training, the remaining fifth fold was used for validation, and the hold-out set was used for test.

This approach presents similar benefits to the leave-one-hospital-out strategy used in previous work[27], with the advantage of a reduced training time. Finally, predictions over the test set were generated by ensembling over the predictions obtained by each of the five trained models. An average pooling voting scheme was applied to each of the model predictions, to produce a single softmax prediction.

	Hospital-wise cross-validation set					
Fold	**0**	**1**	**2**	**3**	**4**	**Hold-out set**
Hospitals validation	STO	GRO, MIL UTR	SFR, VIE	PAR, ZWO ALK, HAG	GOT	AMS
Patients train	646	732	730	728	696	—
Patients validation	237	151	153	155	187	73

Table 2. Distribution of hospitals and patient samples featured in the 5-fold validation sets and hold-out test set.

Metrics

To evaluate the models' voxel-wise performance on the task of residual tumor segmentation, Dice scores were computed between the ground truth annotation and the post-processed binary prediction mask. The Dice scores are reported for the subgroup of patients with residual tumor tissue according to the ground truth annotation, labelled as the 'positive' (P) group, as the Dice score is not stable when applied to empty or nearly empty segmentation masks. The Dice scores for the subgroup of patients with residual tumor according to the ground truth annotation and the network predictions, labelled as the 'true positive' (TP) group, are also reported. Pooled estimates, when computed from each fold's results, are reported for each measurement as mean and standard deviation (indicated by ±) in the tables. For the purpose of assessing the correctness of the predicted tumor size and location, the median absolute volume error (AVE) and the 95% Hausdorff distance (HD95) metrics are also reported.

For the patient-wise classification task of distinguishing patients with gross total resection and patients with residual tumor, a standard sensitivity and specificity approach was conducted represented by the balanced accuracy score (noted bAcc). A residual tumor volume below the clinical volume threshold was thus counted as a negative (i.e., GTR) and as positive otherwise (i.e., RT). Following this consideration, a patient was considered a true positive (TP) if both the ground truth annotation residual tumor volume and detected residual tumor volume were $\geq$ 0.175 ml, for any given Dice score (i.e., $\geq$ 0.01). Conversely, if both volumes were $<$ 0.175 ml, the patient was labelled as a true negative (TN). Patients where the ground truth volume was above the threshold volume and the prediction was below were marked as false negatives (FN), and false positive (FP) vice versa.

In the case of inter-rater variability, the Jaccard score, closely related to the Dice score by $J = \frac{D}{2-D}$, was used to compare the models' performance. The Jaccard was chosen for easy comparison with a previously published work on the same dataset[7].

Statistics

Multiple statistical analyses were carried out to assess and compare the different architectures and input configurations. Statistical tests were conducted on both the cross-validation splits and test set, depending on the task. A significance level of 5% was used throughout the statistical analysis.

For comparing the different input configurations in terms of segmentation performance, Tukey's range tests were performed on the test set, for each of the two architectures. For comparing the two architectures trained on the best input configurations on the segmentation task, a Mann-Whitney U test was conducted on the test set. In terms of classification performance, confidence intervals of individual models were calculated using the bias-corrected and accelerated (BCa) interval method on the test set. For the confidence intervals, significance was determined by assessing whether the intervals overlapped. The models classification performance were also compared using the cross-validation set, computing normal confidence intervals for individual models using pooled estimates from each fold. In the inter-rater study, the Mann-Whitney U test was used to compare each of the architectures with the best input configuration against each individual annotator, as well as the average scores of the two annotator groups novices and experts, and the average over all annotators.

Experiments

The following three experiments were conducted in this study:

(1) Residual tumor segmentation performance study: using the 5-fold cross-validation protocol and segmentation metrics, both nnU-Net and AGU-Net architectures' segmentation performances were compared for the five combinations of input sequences.
(2) Gross total resection classification performance study: using the 5-fold cross-validation protocol, classification metrics, and best input combination identified in the first experiment, both architectures were compared in terms of ability to classify between gross total resection and residual tumor patients.
(3) Inter-rater variability study: the best model from each architecture was benchmarked in terms of segmentation performance against the performance of novice and expert annotators, using the inter-rater variability dataset. For each patient, a consensus agreement annotation has been created using a majority voting approach. Using all eight annotations from both experts and novices, a voxel was defined as belonging to a tumor if annotated by more than half of the annotators. The models' binary predictions and the eight inter-rater annotations were then compared against the ground truth annotations (as used in the hold-out test set) and the consensus annotations.

Results

The studies were performed using multiple machines with the two following specifications: (i) Intel Core Processor (Broadwell, no TSX, IBRS) central processing unit (CPU) with 16 cores, 64GB of RAM, Tesla V100S (32GB) dedicated GPU, and a regular hard-drive and (ii) a GPU server with a total of 256 CPU cores, 2TB of RAM, and six NVIDIA A100-SXM4 (80GB) cards. The AGU-Net architecture was implemented in Python 3.6 with the TensorFlow v1.13.1 library[38]. For the nnU-Net architecture, Python 3.8, PyTorch v1.13.1[39], and the nnU-Net framework v1.7.0[14] were used.

Residual tumor segmentation performance study

Segmentation performances across both architectures, for all input sequences combinations, and only for patients with residual tumor are summarized in Table 3. For both architectures, the lowest average Dice score over the

Input	Prot.	Arch.	Voxel-wise		Patient-wise			
			DSC-P	DSC-TP	Recall	Precision	HD95 (mm)	mAVE (ml)
A	Val	nnU-Net	46.94±24.03	49.51±21.70	94.42±6.69	62.96±6.60	37.55±32.40	0.97
		AGU-Net	37.72±29.54	51.05±22.28	74.10±7.57	80.75±5.79	18.05±17.99	0.76
	Test	nnU-Net	52.38±21.14	53.43±19.77	98.04	70.83	22.99±32.03	1.49
		AGU-Net	38.06±27.45	46.21±22.80	82.35	84.31	21.55±33.40	1.37
B	Val	nnU-Net	52.97±22.66	55.62±19.63	95.08±5.73	66.82±6.06	29.02±31.02	0.51
		AGU-Net	39.71±28.25	51.54±20.59	77.15±7.14	82.30±4.60	16.84±18.09	0.65
	Test	nnU-Net	59.19±20.49	61.61±16.72	96.08	80.65	22.56±33.25	0.57
		AGU-Net	43.76±27.61	53.14±20.23	82.35	87.76	21.28±35.39	0.89
C	Val	nnU-Net	52.43±22.45	54.72±19.77	95.55±7.10	63.70±6.68	35.82±35.02	0.60
		AGU-Net	37.43±28.69	51.09±20.49	73.23±10.68	84.70±3.23	18.76±19.78	0.62
	Test	nnU-Net	58.14±21.01	60.51±17.52	96.08	76.12	25.43±35.27	0.55
		AGU-Net	42.33±27.87	53.97±18.51	78.43	95.24	20.73±36.04	0.55
D	Val	nnU-Net	52.80±22.59	55.26±19.73	95.35±6.40	66.21±5.72	25.08±28.91	0.48
		AGU-Net	41.02±28.08	52.45±20.14	78.28±6.25	85.16±5.24	15.87±16.62	0.55
	Test	nnU-Net	58.05±22.74	60.42±19.61	96.08	79.69	21.23±33.67	0.34
		AGU-Net	40.84±28.62	52.07±20.96	78.43	93.02	19.97±26.64	0.69
E	Val	nnU-Net	53.61±22.57	55.81±19.97	95.86±6.38	63.86±6.93	29.47±31.84	0.64
		AGU-Net	39.44±27.05	48.89±20.92	80.67±5.71	84.58±3.39	16.36±15.42	0.73
	Test	nnU-Net	56.30±21.07	58.60±17.84	96.08	76.12	19.50±30.01	0.69
		AGU-Net	41.23±25.72	47.78±20.93	86.27	89.80	21.38±35.04	1.11

Table 3. Segmentation performances for patients with residual tumor, for both architectures, all input configurations, and over the validation and test sets.

external test set was obtained with configuration A, indicating that solely using T1w-CE MR scans is insufficient for identifying post-operative residual tumor. The addition of the T1w scan as input (i.e., configuration B) provides at least a 5% improvement in Dice scores over the test set for both architectures. This illustrates the additional value of the T1w sequence, presumably due to better distinction between blood and tumor. The inclusion of the FLAIR scan in input configuration C slightly degraded the Dice score compared to input configuration B. Finally, the inclusion of pre-operative data does not seem to improve the performance for any architecture, as the Dice scores for input configuration D are again slightly lower than for configuration B. Further addition of the FLAIR scan in input configuration E leads to a minor decrease in Dice scores compared to configuration D. For both architectures, input configuration B yielded the highest Dice scores on the test set. The highest Dice and true positive Dice scores were obtained with the nnU-Net architecture trained on input configuration B, with respectively 53% and 59% Dice on the validation and test sets. The segmentation performance of nnU-Net trained on config B was significantly better than the segmentation performance of AGU-Net trained on config B (p-value=0.0023, see Table S5, supplementary material). Overall, performances obtained across the test set are stable, in support of generalizability. Likewise, performances over the validation sets from the cross-validation protocol are consistent for input configurations B to E. The same results trends can be observed across both architectures for the true positive Dice, although slightly higher for the positive Dice using the nnU-Net architecture. However, none of the observed differences between input configurations for each architecture were statistically significant according to the Tukey's range tests (see Table S3 and S4, supplementary materials).

Looking at patient-wise performances, models trained with the nnU-Net architecture achieve nearly perfect recall across all configurations for both the validation and test sets. Whereas the patch-wise strategy followed allows for segmenting smaller structures, the loose criterion to consider a network prediction as true positive further strengthens this aspect. Indeed, only a few correctly overlapping voxels between the prediction and the ground truth are needed for residual tumor to be considered satisfactorily identified patient-wise. Due to the full volume approach, models trained with AGU-Net generally struggle to identify small elements, as indicated by an overall around 80% across the board. Conversely, the opposite trend can be noticed in regards to patient-wise precision performance. Models trained with nnU-Net tend to perform more erroneous predictions as indicated by average precision scores below 70%, whereas AGU-Net models tend to be more precise with precision scores up to 95%. Similarly, the HD95 for nnU-Net evaluated on the validation set are about twice the distances of the AGU-Net, most likely due to the high rate of false positives produced by the patch-wise approach. This effect is considerably reduced for the test set, probably due to the effect of ensembling of the five models from cross-validation on reducing the number of false positives. The median AVE are quite similar between the models, and acceptable given the average volumes on the two datasets of 3.06 ml and 1.93 ml on the validation and test sets, respectively.

From the segmentation performances analysis, the best results have been obtained with the nnU-Net architecture using input configuration B. Visual comparisons are provided in Fig. 3 between the two architectures using the best input configuration for some patients from the test set, one featured per row. In the top row, both models

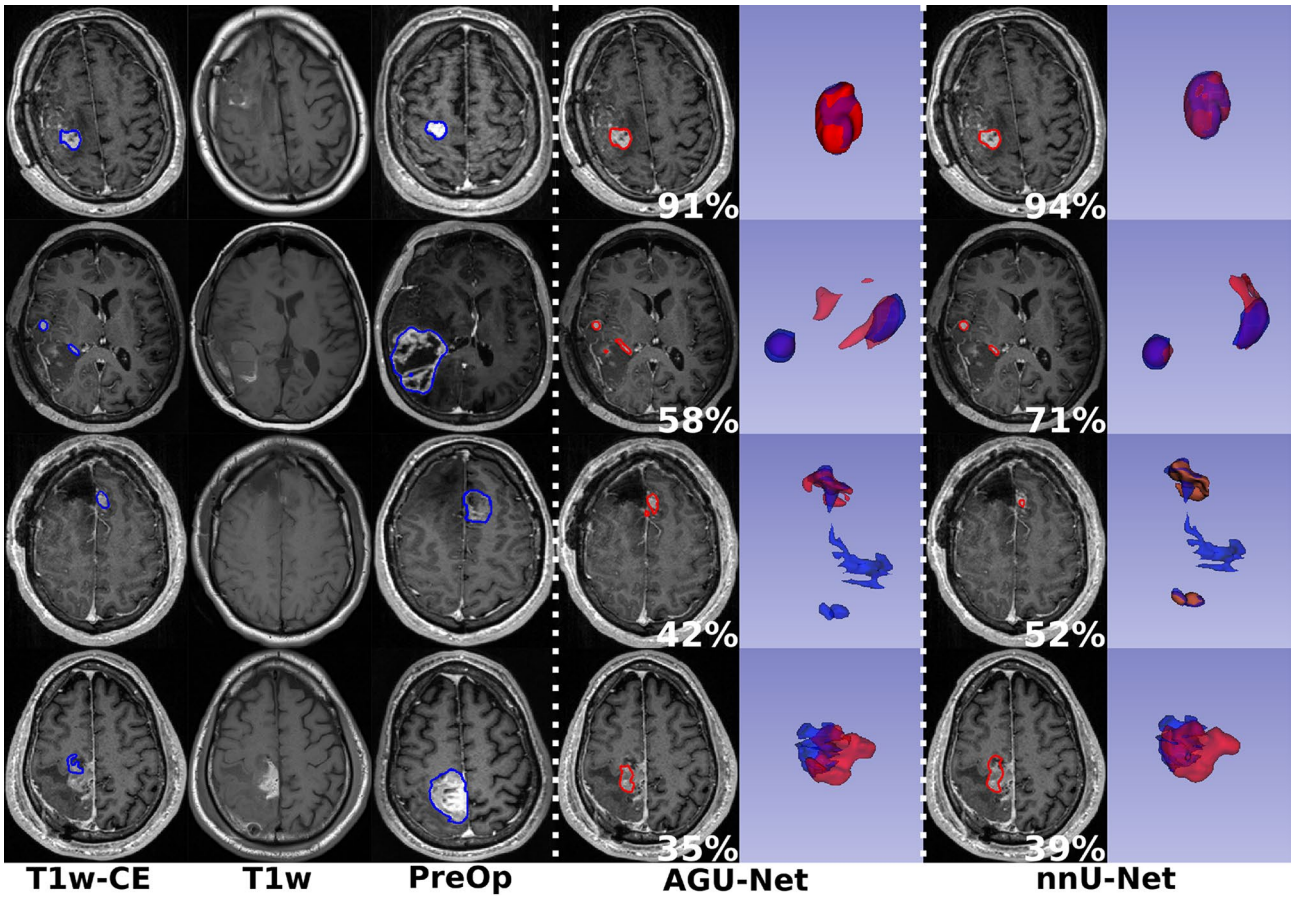

Figure 3. Segmentation comparison between the manual ground truth (in blue) and the binary predictions (in red) for the two architectures using configuration B, over the test set. One patient is featured per row, the patient-wise Dice is reported in white, and a 3D representation of the overlap is included (best viewed digitally and in color).

achieved excellent segmentation with a Dice score above 90%. In the second row, a multifocal post-operative residual tumor case is featured whereby the AGU-Net model produced one false positive component as can be seen in red in the 3D representation. For the third row, a challenging multifocal and fragmented residual tumor case is displayed where both models failed to segment the largest component. Finally, in the last row, oversegmentation was performed using both models leading to Dice scores below 40%.

Gross total resection classification performance study

Classification performances between patients with residual tumor and gross total resections, across both architectures and for all input configurations, are reported in Table 4. The first noticeable result is the overall tendency of the nnU-Net architecture to oversegment, resulting in a perfect recall over both the test set and validation set, for a really poor specificity often below 30%. Overall, nnU-Net achieves balanced accuracy scores barely above 0.5 for all input configurations, which means the classification performance is only slightly better compared to the average score of random guessing (i.e., 0.5). Conversely, models trained with the AGU-Net architecture are more conservative leading to higher specificity scores, up to 90% for input configuration C, and reasonably high recall/sensitivity values above 80%. The classification performances of AGU-Net configurations B-E were all significantly better than all of the nnU-Net configurations, evaluated on the cross-validation set (see Table S7, supplementary materials). However, the difference was only significant between AGU-Net config B-E and nnU-Net config A when evaluated on the test set. In contrast to segmentation performances, the successive addition of MR scans within the input configuration lead to improved classification performances for both architectures, although none of these improvements were statistically significant (see Tables S6 and S7, supplementary materials). One apparent difference is the added value of the FLAIR sequence with the AGU-Net architecture, further increasing the specificity and balanced accuracy, unlike performances with the nnU-Net architecture.

From the classification performance analysis, the best results on the test set according to the balanced accuracy have been obtained with the AGU-Net architecture using input configuration C. However, the best results on the validation sets are achieved with input configuration E. In a clinical scenario, a high sensitivity has higher priority than a high specificity, as long as the trade-off is reasonable. AGU-Net trained with input configuration E

Exp.	Data	Arch.	Patient-wise		
			Sensitivity	Specificity	bAcc
A	Val	nnU-Net	99.81±0.35	2.53±2.21	51.17±1.22
		AGU-Net	79.70±6.69	68.01±10.43	73.86±4.94
	Test	nnU-Net	100.00	4.55	52.27
		AGU-Net	84.31	63.64	73.98
B	Val	nnU-Net	99.47±0.71	18.04±4.41	58.75±2.30
		AGU-Net	81.25±6.47	71.01±5.36	76.13±4.12
	Test	nnU-Net	98.04	45.45	71.75
		AGU-Net	84.31	72.73	78.52
C	Val	nnU-Net	99.81±0.35	5.64±3.44	52.73±1.76
		AGU-Net	79.29±10.08	74.00±11.13	76.64±4.87
	Test	nnU-Net	100.00	27.27	63.64
		AGU-Net	78.43	90.91	84.67
D	Val	nnU-Net	99.66±0.44	15.28±6.85	57.47±3.55
		AGU-Net	82.80±5.27	73.00±14.63	77.90±6.44
	Test	nnU-Net	100.00	40.91	70.45
		AGU-Net	78.43	86.36	82.40
E	Val	nnU-Net	100.00	6.12±4.30	53.06±2.15
		AGU-Net	85.61±4.83	72.63±9.39	79.12±4.60
	Test	nnU-Net	100.00	27.27	63.64
		AGU-Net	86.27	77.27	81.77

Table 4. Gross total resection versus residual tumor classification performances for both architectures, all input configurations, and over the validation and test sets.

is therefore the preferred model for classification. This configuration achieves the highest sensitivity for all input configurations while still achieving a reasonable specificity, higher than configurations A and B.

Inter-rater variability study

For the 20 patients constituting the inter-rater variability dataset, a comparison of the Jaccard scores, obtained between each rater and the best model from each architecture, are reported in Fig. 4. The two references used in the analysis are also evaluated against each other to illustrate how they differ. The scores differ slightly between the two considered references since the reporting over "positive" cases (i.e., with residual tumor) is inherently

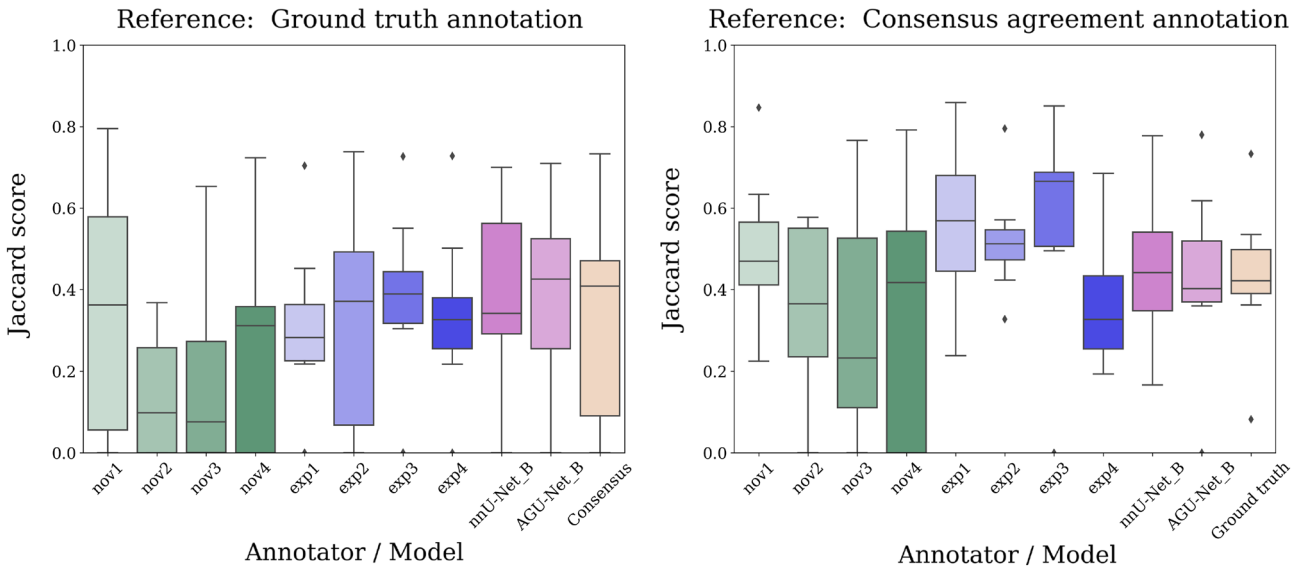

Figure 4. Inter-rater Jaccard score variability over a subset of the AMS cohort. To the left, the ground truth annotation used for training served as segmentation of reference. To the right, the reference segmentation was a consensus agreement between annotations from all raters.

linked to the segmentation used as reference. As all configurations B-E yielded very similar segmentation performance scores, with no significant differences between configurations for each architecture (see Table S3 and S4, supplementary materials), input configuration B was selected as the best configuration for both architectures. In addition, configuration B also has the advantage of minimal requirements in terms of input sequences. Using the ground truth annotation from the test set as a reference segmentation, both architectures achieved Jaccard scores within the variability range of the novice and expert annotators, and the nnU-Net architecture even demonstrated significantly better performance compared to two of the novice annotators. However, no significant differences were observed between any of the two architectures and the novice and expert annotators on a group level (see Table S8 and S9, supplementary material). Using the consensus agreement annotation as the reference segmentation, the AGU-Net model achieved slightly poorer Jaccard scores than the majority of the expert human raters, however only one of the expert annotators scores were significantly better than the model (see Table S6, supplementary materials), and the model remained within the variability of all annotators on a group level. The nnU-Net model achieved scores similar to the variability range of the expert raters, also when compared to the consensus agreement annotation, and none of the differences with the annotators were significant.

Discussion

In this multicenter study, the feasibility of post-operative residual tumor segmentation with deep neural networks was assessed. Two state-of-the-art architectures for pre-operative glioblastoma segmentation were compared: nnU-Net and AGU-Net. Both architectures were trained on five different combinations of early post-operative and pre-operative MR scans as input, and benchmarked in terms of segmentation and classification performances compared with manual rating. The main finding is that automatic segmentation performances are comparable to human rater performance on real world MRI scans, requiring early post-operative T1w-CE and T1w MRI scans only. In addition, the trained automated models have shown promising ability to classify patients who underwent gross total resection from patients exhibiting post-operative residual tumor.

The multimodal and multicentric dataset in this study is the largest cohort used for the task of early post-operative glioblastoma segmentation, with a total of 956 patients. Regarding the dataset curation, our strict inclusion criteria required availability of all four MR scans as input (i.e., post-operative T1w-CE, T1w, FLAIR, and pre-operative T1w-CE) for each patient. Whereas this decision was motivated by a simpler method design, approximately 150 patients were excluded as one or more MR scans were missing. A relaxation of the inclusion criteria would increase the size of the dataset, and open the possibility to generate a more diverse set of input MR scans, including for example T2-weighted images. Ideally, the trained methods should be able to deal with a sparse set of inputs, where one or more MR scans are missing. The trained models should be used off the shelf, by replacing missing sequences with empty volumes, synthetically generated sequences, or allowing missing inputs using sparsified learning techniques[40].

In their ability to segment post-operative tumor, nnU-Net and AGU-Net exhibit strengths and weaknesses inherent to their design. Through a patch-wise approach, nnU-Net models are able to segment relatively small structures, having access to more fine-grained details from the use of MR scans close to their initial resolution. Considering the relatively small volumes and fragmented state for residual tumors, nnU-Net models are able to achievesignificantly higher Dice score and recall performances than the AGU-Net. On the other hand, models trained using the AGU-Net approach are following a full volume approach, largely downsampling the input MR scans, hindering the ability for detecting smaller structures. However, such models appear to be more conservative in their predictions, hence heavily reducing the amount of false positives enabling to reach high specificity performances, and significantly higher balanced accuracy scores than the nnU-Net architecture for some configurations. Regarding the different input configurations, the biggest impact on segmentation performances comes from combining EPMR T1w-CE and T1w scans, which corresponds to the favored approach as well in clinical practice. The inclusion of additional MR sequences seems to add little to segmentation performances, which is confirmed by the absence of statistical significant differences between input configurations for each of the models. Adapting the convolution blocks, filter sizes, or other elements of the architectures might be needed for letting the number of trainable parameters to evolve according to the number of inputs, instead of a fixed amount of parameters.

The validation studies described in this article served the two purposes of investigating the predictive ability and capacity to generalize of the trained models. This is obtained through the use of a unique test set, and equally distributed hospital-stratified validation sets. Our selection for a specific hold-out hospital as a test set was based on the availability of manual annotations from multiple raters, allowing to perform, in addition, an inter-rater variability study. Regarding the computation of the reported metrics, the rationale for only including the true positive patients in the segmentation performances lies in the Dice score computation itself. Indeed, cases with a GTR preclude calculation of a Dice score. Therefore, the validation studies include a separate experiment to classify patients into those with a GTR and those with residual tumor.

The inter-rater variability study demonstrated that residual tumor segmentation performance is on par with the average human expert annotator performance, when evaluated against an independent ground truth segmentation. Even when evaluated against the consensus agreement annotation, which is by definition biased towards each of the human annotators included in the study, the best segmentation model achieves scores similar to the individual expert annotations. The consensus agreement annotation based on a majority voting scheme over all annotations from the eight different annotators should be considered the gold standard for defining the residual tumor. However, this is not achievable in a real-world clinical scenario, where even an exact delineation of the tumor remnant from one human annotator is rarely performed. The best available segmentation models for pre-operative segmentation achieve Dice scores of up to 90%, but the inter-rater variability study shows that this is far from realistic in the early post-operative case. Indeed, the particular challenges of early post-operative

segmentation, such as small and fragmented tumors with an average tumor volume of only 3 ml in this dataset, in contrast to the pre-operative case with an average tumor volume of 35 ml, makes it difficult even for human expert annotators to achieve Dice scores of more than 60%.

The proposed automatic method for residual tumor segmentation should thus be considered an acceptable alternative to the current standard practice for evaluating the tumor remnant after surgery, as the average performance of the method lies within the variability range of individual expert annotators. Such segmentation performances are even achieved with the exclusive use of post-operative MR sequences as model inputs (T1w-CE, T1w, and FLAIR), whereas the addition of pre-operative information (pre-operative T1w-CE and label) retains the model performance on similar levels. Thus, in clinical practice, our trained models could be deployed even in the absence of pre-operative scans, as long as at least the T1w-CE and T1w post-operative sequences are available, to establish an automated and relatively fast method for the segmentation task. On a second level, the output segmentation masks can be used to differentiate between patients with remnant tumor after surgery and gross total resection patients, with increasing balanced accuracy performance as more sequences are added to the model inputs. Our early post-operative glioblastoma segmentation models have been made freely available in the Raidionics environment[41].

In spite of promising reported performances, the task of early post-operative glioblastoma segmentation is far from accomplished. The full extent of residual tumor, often very fragmented around the resection cavity, is never wholly captured. In future work, the pre-operative MR scans and tumor location should be better leveraged as the residual tumor is bound to lie in its vicinity. Focusing the search solely within a region of interest might help retaining a higher image resolution, for better segmentation of small structures. Nevertheless, competitive pre- and post-operative glioblastoma segmentation models are now publicly available, opening the door to clinically-oriented validation studies. Assuming a positive outcome, the use of automatic models and methods would be highly beneficial in a clinical setting to collect parameters currently obtained through eyeballing or diameter estimation, hence yielding reproducible and deterministic significance.

Conclusion

In this study, two state-of-the-art neural network architectures for glioblastoma segmentation were trained and thoroughly validated on a large cohort of 956 patients. Automatic segmentation performances are on par with human rater performance on real world MRI scans, requiring early post-operative T1w-CE and T1w MRI scans only. In addition, the presented models have shown promising readiness for automatically distinguishing between patients who underwent gross total resection, and patients with residual tumor. The prognostic value of the automated method should be assessed in future studies.

Data availability

The dataset analysed in this study is available from the corresponding author on reasonable request. The best trained models along with source code for validation and inference are made publicly available, and the accession codes can be found under 'Additional Information'. The best trained AGU-Net model can be accessed at https://github.com/raidionics/Raidionics-models/releases/tag/1.2.0. The best trained nnU-Net model can be accessed at https://gitlab.com/picture-production/picture-nnunet-package/tree/0.3.7. The source code used for computing the metrics can be accessed at https://github.com/dbouget/validation_metrics_computation. Inference on new patients can be performed using the Raidionics software which is openly available at https://github.com/raidionics/Raidionics.

Received: 16 May 2023; Accepted: 19 October 2023
Published online: 02 November 2023

References

1. Davis, M. E. Glioblastoma: Overview of disease and treatment. *Clin. J. Oncol. Nurs.* **20**, 1–8. https://doi.org/10.1188/16.CJON.S1.2-8 (2016).
2. 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-Oncol. Adv.* **3**, 1–12. https://doi.org/10.1093/noajnl/vdab008 (2021).
3. Coburger, J., Wirtz, C. R. & König, R. W. 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. *J. Neurosurg. Sci.* **61**, 233–244. https://doi.org/10.23736/S0390-5616.16.03284-7 (2017).
4. Aabedi, A. A. *et al.* Association of neurological impairment on the relative benefit of maximal extent of resection in chemoradiation-treated newly diagnosed isocitrate dehydrogenase wild-type glioblastoma. *Neurosurgery* **90**, 124–130. https://doi.org/10.1227/NEU.0000000000001753 (2022).
5. Berntsen, E. M. *et al.* Volumetric segmentation of glioblastoma progression compared to bidimensional products and clinical radiological reports. *Acta Neurochir.* **162**, 379–387. https://doi.org/10.1007/s00701-019-04110-0 (2020).
6. Wen, P. Y. *et al.* Updated response assessment criteria for high-grade gliomas: Response assessment. *Neuro-Oncol. Work. Group*https://doi.org/10.1200/JCO.2009.26.3541 *(2010).*
7. Visser, M. *et al.* Inter-rater agreement in glioma segmentations on longitudinal MRI. *NeuroImage Clin.* **22**, 101727. https://doi.org/10.1016/j.nicl.2019.101727 (2019).
8. Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imag.* **34**, 1993–2024. https://doi.org/10.1109/TMI.2014.2377694 (2015).
9. Shelhamer, E., Long, J. & Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 640–651. https://doi.org/10.1109/TPAMI.2016.2572683 (2017).
10. Wang, R. *et al.* Medical image segmentation using deep learning: A survey. *IET Image Proc.* **16**, 1243–1267 (2022).
11. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes Comput. Sci.*https://doi.org/10.1007/978-3-319-24574-4_28 *(2015).*
12. Kamnitsas, K. *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78. https://doi.org/10.1016/j.media.2016.10.004 (2017).

13. Schlemper, J. *et al.* Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **53**, 197–207. https://doi.org/10.1016/j.media.2019.01.012 (2019).
14. 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. *Nat. Methods* **18**, 203–211. https://doi.org/10.1038/s41592-020-01008-z (2021).
15. Luu, H. M. & Park, S.H. Extending nn-unet for brain tumor segmentation. In Crimi, A. & Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 173–186, https://doi.org/10.1007/978-3-031-09002-8_16(Springer International Publishing, Cham, 2022).
16. 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 (2022).
17. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66. https://doi.org/10.1109/TSMC.1979.4310076 (1979).
18. Cordova, J. S. *et al.* Quantitative tumor segmentation for evaluation of extent of glioblastoma resection to facilitate multisite clinical trials. *Trans. Oncol.* **7**, 40–47. https://doi.org/10.1593/tlo.13835 (2014).
19. Odland, A. *et al.* Volumetric glioma quantification: Comparison of manual and semi-automatic tumor segmentation for the quantification of tumor growth. *Acta Radiol.* **56**, 1396–1403. https://doi.org/10.1177/0284185114554822 (2015).
20. Zeng, K. et al. Segmentation of Gliomas in Pre-operative and Post-operative Multimodal Magnetic Resonance Imaging Volumes Based on a Hybrid Generative-Discriminative Framework BT—Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 184–194, https://doi.org/10.1007/978-3-319-55524-9_18(Springer International Publishing, Cham, 2016)
21. Chow, D. S. *et al.* Semiautomated volumetric measurement on postcontrast MR imaging for analysis of recurrent and residual disease in glioblastoma multiforme. *Am. J. Neuroradiol.* **35**, 498–503. https://doi.org/10.3174/ajnr.A3724 (2014).
22. Krivoshapkin, A. L. *et al.* Automated Volumetric Analysis of Postoperative Magnetic Resonance Imaging Predicts Survival in Patients with Glioblastoma. *World Neurosurg.* **126**, e1510–e1517. https://doi.org/10.1016/j.wneu.2019.03.142 (2019).
23. Zhu, Y. *et al.* Semi-automatic segmentation software for quantitative clinical brain glioblastoma evaluation. *Acad. Radiol.* **19**, 977–985. https://doi.org/10.1016/j.acra.2012.03.026 (2012).
24. Dhara, A. K. et al. Interactive Segmentation of Glioblastoma for Post-surgical Treatment Follow-up. Proceedings - International Conference on Pattern Recognition 2018-August, pp 1199–1204, https://doi.org/10.1109/ICPR.2018.8545105(2018).
25. Meier, R. *et al.* Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma. *J. Neurosurg.* **127**, 798–806. https://doi.org/10.3171/2016.9.JNS16146 (2017).
26. Ghaffari, M. *et al.* Automated post-operative brain tumour segmentation: A deep learning model based on transfer learning from pre-operative images. *Magn. Reson. Imag.* **86**, 28–36. https://doi.org/10.1016/j.mri.2021.10.012 (2022).
27. Bouget, D. *et al.* Glioblastoma surgery imaging-reporting and data system: Validation and performance of the automated segmentation task. *Cancers*https://doi.org/10.3390/cancers13184674 *(2021).*
28. Bouget, D., Pedersen, A., Hosainey, S. A. M., Solheim, O. & Reinertsen, I. Meningioma segmentation in T1-weighted MRI leveraging global context and attention mechanisms. *Front. Radiol.*https://doi.org/10.3389/fradi.2021.711514 *(2021).*
29. Bouget, D. *et al.* Preoperative brain tumor imaging: Models and software for segmentation and standardized reporting. *Front. Neurol.*https://doi.org/10.3389/fneur.2022.932219 *(2022).*
30. Kommers, I. *et al.* Glioblastoma surgery imaging-reporting and data system: Standardized reporting of tumor volume, location, and resectability based on automated segmentations. *Cancers* **13**, 2854. https://doi.org/10.3390/cancers (2021).
31. Nabors, L. B. *et al.* Central nervous system cancers, version 1.2017 featured updates to the NCCN guidelines. *JNCCN J. Nat. Compr. Cancer Netw.* **15**, 1331–1345. https://doi.org/10.6004/jnccn.2017.0166 (2017).
32. Garcia-Ruiz, A. *et al.* Precise enhancement quantification in post-operative MRI as an indicator of residual tumor impact is associated with survival in patients with glioblastoma. *Sci. Rep.* **11**, 1–10. https://doi.org/10.1038/s41598-020-79829-3 (2021).
33. Stupp, R., Brada, M., van den Bent, M. J., Tonn, J. C. & Pentheroudakis, G. High-grade glioma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **25**, 93–101. https://doi.org/10.1093/annonc/mdu050 (2014).
34. Stummer, W. *et al.* Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: A randomised controlled multicentre phase III trial. *Lancet. Oncol.* **7**, 392–401. https://doi.org/10.1016/S1470-2045(06) (2006).
35. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41. https://doi.org/10.1016/j.media.2007.06.004 (2008).
36. Avants, B. B. *et al.* Advanced normalization tools (ants). *Insight J.* **2**, 1–35 (2009).
37. Pedersen, A. & Bouget, D. andreped/gradientaccumulator: v0.3.1, https://doi.org/10.5281/zenodo.7582309(2023).
38. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (2015).
39. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, 8024–8035 (Curran Associates, Inc., 2019).
40. Grøvik, E. *et al.* Handling missing MRI sequences in deep learning segmentation of brain metastases: A multicenter study. *NPJ Digital Med.* **4**, 33. https://doi.org/10.1038/s41746-021-00398-4 (2021).
41. Bouget, D. *et al.* Preoperative brain tumor imaging: Models and software for segmentation and standardized reporting. *Front. Neurol.*https://doi.org/10.3389/fneur.2022.932219 *(2022).*

Acknowledgements

Data were processed in digital labs at HUNT Cloud, Norwegian University of Science and Technology, Trondheim, Norway. FB is supported by the National Institute for Health Research (NIHR) biomedical research centre at UCLH. The PICTURE project is sponsored by 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 and the Anita Veldman foundation, CCA2018-2-17. R.H.H. is supported by a grant from The Research Council of Norway, grant number 323339. D.B., I.R., and O.S. are partly funded by the Norwegian National Research Center for Minimally Invasive and Image-Guided Diagnostics and Therapy.

Author contributions

Conceptualization: R.H.H., A.F., D.B., R.E., I.R., O.S., and P.C.D.W.H.; Data curation: I.K., H.A., F.B., L.B., M.B., T.D., M.N., J.F., S.H.J., A.I., B.K., R.T., E.M., D.M., P.R., M.R., L.S., T.S., T.A., M.W., G.W., M.W., A.Z., P.M., A.J., O.S., P.C.D.W.H.; Methodology, Investigation, Formal analysis, and Validation: R.H.H., A.F., D.B., and R.E.;

Funding acquisition and project administration: I.R., O.S, and P.C.D.W.H.; Software: D.B., A.P., R.H.H.; Writing - original draft: R.H.H., A.F., D.B., A.P., R.E., I.R., O.S., and P.C.D.W.H.; Writing - review & editing: all authors.

Funding

Open access funding provided by Norwegian University of Science and Technology.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-45456-x.

Correspondence and requests for materials should be addressed to R.H.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023