Computers in Biology and Medicine 142 (2022) 105237

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Segmenting pediatric optic pathway gliomas from MRI using deep learning

Jakub Nalepa [a,b,*], Szymon Adamski [b], Krzysztof Kotowski [b], Sylwia Chelstowska [c], Magdalena Machnikowska-Sokolowska [d], Oskar Bozek [d], Agata Wisz [d], Elzbieta Jurkiewicz [c]

[a] Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland
[b] Graylight Imaging, Gliwice, Poland
[c] Children's Memorial Health Institute, Warsaw, Poland
[d] Division of Diagnostic Imaging, Department of Radiology and Nuclear Medicine, Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland

ARTICLE INFO

Keywords:
Optic pathway glioma
Deep learning
Pre-training
Transfer learning
Segmentation
LGG
HGG

ABSTRACT

Optic pathway gliomas are low-grade neoplastic lesions that account for approximately 3–5% of brain tumors in children. Assessing tumor burden from magnetic resonance imaging (MRI) plays a central role in its efficient management, yet it is a challenging and human-dependent task due to the difficult and error-prone process of manual segmentation of such lesions, as they can easily manifest different location and appearance characteristics. In this paper, we tackle this issue and propose a fully-automatic and reproducible deep learning algorithm built upon the recent advances in the field which is capable of detecting and segmenting optical pathway gliomas from MRI. The proposed training strategies help us elaborate well-generalizing deep models even in the case of limited ground-truth MRIs presenting example optic pathway gliomas. The rigorous experimental study, performed over two clinical datasets of 22 and 51 multi-modal MRIs acquired for 22 and 51 patients with optical pathway gliomas, and a public dataset of 494 pre-surgery low-/high-grade glioma patients (corresponding to 494 multi-modal MRIs), and involving quantitative, qualitative and statistical analysis revealed that the suggested technique can not only effectively delineate optic pathway gliomas, but can also be applied for detecting other brain tumors. The experiments indicate high agreement between automatically calculated and ground-truth volumetric measurements of the tumors and very fast operation of the proposed approach, both of which can increase the clinical utility of the suggested software tool. Finally, our deep architectures have been made open-sourced to ensure full reproducibility of the method over other MRI data.

1. Introduction

Optic pathway gliomas (OPGs) are low-grade tumors accounting for approximately 3–5% of pediatric brain tumors [1,2]. These neoplasms may arise sporadically or in association with neurofibromatosis type 1 (NF1), and the incidence of OPGs is significantly higher in patients with NF1 [3]—it was shown that 15–20% of children with NF1 will develop an optic pathway tumor [4]. In the majority of cases, NF1-associated OPGs are classified as pilocytic astrocytomas, whereas the symptoms depend on their specific locations [5]. Interestingly, 45–65% of OPGs involve only the prechiasmatic pathway, while 25%–50% involve the chiasmatic region with or without the optic nerves; the optic tracts are affected (with or without the chiasmatic and/or prechiasmatic regions) in the remaining cases [6]. Although they are low grade gliomas (WHO 1) with a good prognosis in general [3], they show a high complexity and variability in shape, texture, image enhancement, and degrees of cystic change [1,7], and they may cause severe complications like vision loss.

Tumor management in OPGs includes radiological observation and administering chemo- or radiotherapy if the progression is fast [8]. An appropriate patient's monitoring and selection of therapy are the most disputable aspects of this disease [1,9]. The longitudinal evaluation of tumor size is a key diagnostic procedure when tracking the progression and efficacy of the treatment. It is usually based on manual linear measurements performed on a single slice [10], that are subjective, hard to reproduce and often inaccurate. To tackle these issues, several automated segmentation frameworks for longitudinal analysis of tumors have been introduced. They allow us to quantify the tumor volume in the repeatable and accurate way, hence make the entire process independent from human errors and not suffering from the lack of

* Corresponding author. Department of Algorithmics and Software, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland.
E-mail address: jnalepa@ieee.org (J. Nalepa).

https://doi.org/10.1016/j.compbiomed.2022.105237
Received 21 October 2021; Received in revised form 13 January 2022; Accepted 14 January 2022
Available online 17 January 2022

Table 1
The summary of the algorithms for automatic segmentation of pediatric OPG from MRI. We present the underlying approach used in the algorithm, the MRI sequences utilized in the work, the size of the dataset used in the experiments (quantified as the number of MRIs, #MRIs, and the corresponding number of patients, pts), and the age of patients enrolled into the study (for Artzi et al. [25] and our work, we report the average age and its standard deviation in the parenthesis—this information is not available for other works).

Ref.	Year	Algorithm	MRI sequences	#MRIs	Age (yrs)
[26]	2010	Atlas probabilistic model	T1, T2, FLAIR	15 MRIs (15 pts)	3–7
[27]	2011	Atlas probabilistic model	T1, T2, FLAIR	24 MRIs (3 pts)	3.5–4
[28]	2012	Atlas probabilistic model + spectral angle mapper [29]	T1, T2, FLAIR	28 MRIs (7 pts)	2–7
[30]	2017	Shape and appearance feature extractors + DNN	T1c, T2, FLAIR	20 MRIs (20 pts[a])	N/A
[25]	2020	U-Net + ResNet and fuzzy *c*-means clustering	T1c, T2	202 MRIs (29 pts)	5.7 ± 5.4
Ours	2021	nnU-Nets	T2, FLAIR	22 MRIs (22 pts)	2 – 14 (7.5 ± 3.5)
				51 MRIs (51 pts)	3 – 19 (9.0 ± 4.1)

[a] Only 10 patients had diagnosed OPGs.

reproducibility, that are inherently related to manual segmentation [11]. In this paper, we follow this research pathway, and propose a deep learning algorithm for OPG segmentation from multi-modal MRI—we summarize our contributions in Section 1.2.

1.1. Related literature

The majority of brain tumor segmentation algorithms in the literature focus on glioblastoma (former name glioblastoma multiforme, GBM) in multi-sequence MRI [12–14] popularized by the BraTS (Brain Tumor Segmentation) challenge [15,16]. These algorithms can be divided into three main categories [17,18]—atlas-, image analysis-, and machine learning-based techniques. The atlas-based models segment the scans by comparing them to the manually curated reference images (referred to as *atlases*) representing the natural anatomical variability of the brain tissue [19]. Classical image analysis algorithms usually classify the voxels based on their intensity (through employing various thresholding approaches) or the appearance of their neighborhood (utilizing the texture- or shape-based filters, or region-growing algorithms). Machine learning approaches are split into the conventional techniques which require manual feature engineering, including feature extraction commonly followed by feature selection which aims at selecting a subset of the most discriminative image features [20,21], and deep learning algorithms which learn features from the data automatically during the training process [22]. Currently, deep learning is the state-of-the-art technique which won the majority of recent biomedical segmentation competitions [23,24].

Since OPGs have vastly different locations and characteristics than GBMs, it may not be possible to effectively use GBM-trained models for them (we also tackle this issue in the work reported here, and experimentally verify it in Section 3). There exist several techniques for automatic segmentation and classification of OPGs in children. They use different sets of MRI sequences including pre- and post-contrast T1 (T1c), T2, and FLAIR, but only T2 is common to all the methods (Table 1). The first reported OPG segmentation techniques used atlas-guided probabilistic methods tested on small datasets containing several 2–7 year old children [26–28]. These algorithms addressed the problem of internal classification of OPG into solid, enhancing, and cystic components which ultimately increases the clinical usefulness of the results, especially in longitudinal studies (the change of the component type does not affect the total tumor volume but may be crucial diagnostic information). However, besides small sample sizes, these methods had several limitations, including lack of reliable brain atlases for children, deficient performance for OPGs extending beyond the chiasm area, and volume under-estimation [28]. The first application of a deep neural network for segmenting OPGs was reported by Mansoor et al. [30]—here, a classifier exploiting shape and appearance features of the automatically segmented anterior visual pathways (AVPs) was used, so it was strictly limited to the subset of OPGs located directly on the AVP. In one of the latest works [25], the authors utilized the U-Net architecture [31] initialized with the ResNet-34 model weights pre-trained on ImageNet data, encompassing thousands of non-medical RGB images [32], and fine-tuned the model using three-channel images containing a post-contrast T1 axial slice (channel 1), a T2 axial slice (channel 2), and a black image (channel 3). This approach tries to mitigate the problem of small OPG datasets, but on the other hand it involves many redundant connections and heavy domain shift from ordinary RGB images to multi-sequence MRI slices. Additionally, the OPG dataset containing more than 200 MRI studies used in

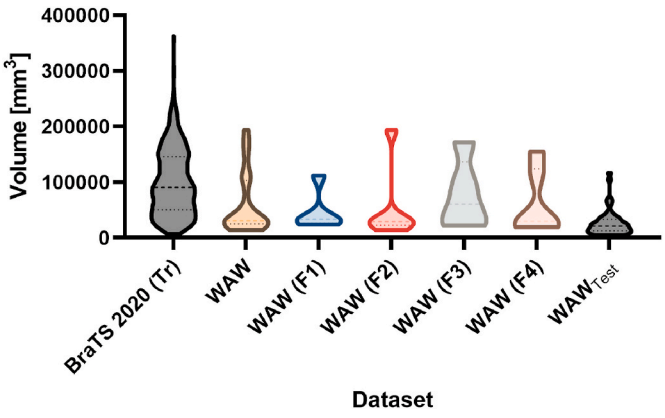

Fig. 1. Distribution of the whole tumor volume (in mm^3) in BraTS 2020 (Tr), WAW (without and with division into non-overlapping folds, referred to as F1–F4), and in WAW_{Test}.

Table 2
Descriptive statistics of the whole tumor volume (mm^3) in BraTS 2020 (Tr), WAW (without and with division into non-overlapping folds, referred to as F1–F4), and in WAW_{Test}.

Dataset	Pts	Min.	25p	Median	75p	Max.	Mean	St. dev.
BraTS 2020 (Tr)	369	7285	50 459	90 740	145 809	361 783	99 547	59 429
WAW	22	14 156	24 890	30 163	103 423	193 419	59 213	55 155
WAW_{Test}	51	4661	12 234	20 870	33 339	116 458	26 409	22 815
WAW (F1)	5	23 699	26 459	33 383	72 828	111 914	46 392	36 852
WAW (F2)	6	14 156	22 504	29 166	72 530	193 419	53 905	68 642
WAW (F3)	5	22 229	23 774	60 337	136 186	171 780	76 051	62 217
WAW (F4)	6	19 550	19 552	29 722	123 919	154 906	61 174	58 270

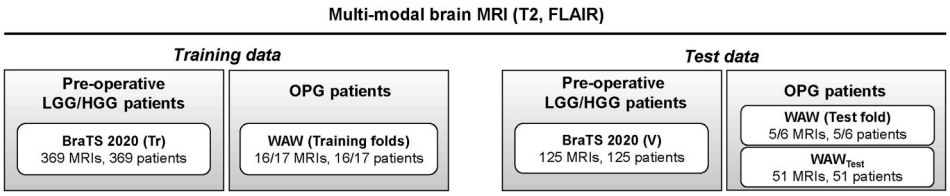

Fig. 2. Training and test MRIs used in this study.

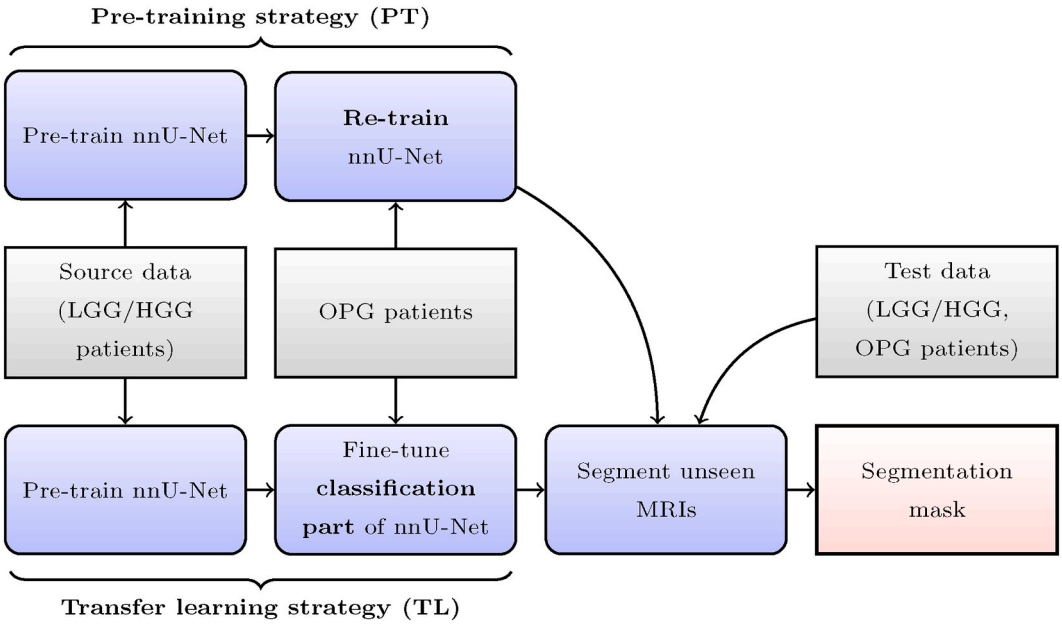

Fig. 3. We utilize two training strategies (pre-training and transfer learning) to effectively deal with the limited OPG ground-truth datasets (in this study, we also investigate the performance of nnU-Nets trained exclusively on the BraTS/WAW cohort). In both strategies, we benefit from the available training scans of the same modality (T2 and FLAIR MRI), but in a slightly different task (LGG/HGG vs. OPG detection and segmentation). For both strategies, we boldface the step that affects the pre-trained nnU-Net (either all weights in PT, or a small subset of all trainable weights of the classification part of the deep architecture in TL).

this approach is not publicly available, so it is impossible to reproduce or compare the results over the very same data.

The performance achieved by the automated OPG segmentation techniques is commonly quantified by a mean volume overlap using the DICE coefficient that ranges from 0.694 ± 0.088 [28], up to 0.761 ± 0.011 [25] (or 0.770 ± 0.180 considering only OPGs located directly on the AVP [30]), depending on the test datasets and validation approaches. This is significantly lower than the performance commonly reported for the GBM whole-tumor segmentation, and reaching the DICE of 0.890 in the latest BraTS 2020 [23]. It is worth mentioning that the datasets for OPG are much smaller and less explored, hence elaborating well-generating large capacity learners becomes significantly more challenging. Other metrics used for assessing the segmentation quality include the mean surface distance error (amounting to 0.737 ± 0.319 mm in Ref. [28]), and the Hausdorff distance: 6.94 ± 6.25 mm in Ref. [30]. In Table 1, we gather a concise summary of the existing techniques, and highlight the method introduced in this work.

Overall, although there exist the algorithms for automated OPG segmentation, it is challenging to deploy them in practice, due to the very limited amount of ground-truth OPG delineations that could be used for training well-generalizing models. We approach this issue through building upon the recent architectural advances in the field, and we benefit from the transfer learning and pre-training strategies (that are independent from the underlying architecture), which allow us to elaborate the algorithms that can accurately segment OPG from MRI, even if the number of ground-truth OPG samples is small.

1.2. Contribution

In this work, we tackle the problem of automatic segmentation of pediatric OPGs from MRI using deep learning. Our contributions are multi-fold and can be summarized by the following bullet points:

- We introduce a deep learning algorithm for segmenting OPGs from MRI scans, together with the strategies that allow us to train the deep learning models over very limited OPG MRI data while still maintaining high generalization capabilities (Section 2.2). We build upon the nnU-Net framework which established the current state of the art in the low- and high-grade glioma (LGG/HGG) detection and segmentation from MRI, consistently outperforming other deep learning and classical algorithms for this task (the nnU-Net segmentation engine took the first place in the BraTS 2020 competition [33]).
- We investigate the capabilities of the algorithms in the rigorous multi-fold experimental study (Section 3), involving 13 deep learning models trained and validated over a public domain database of multi-modal MRI images captured for pre-operative LGG/HGG patients: 293 HGG and 76 LGG patients with known ground-truth segmentations constituting the training set BraTS (Tr), and 125 patients without manual delineations, constituting the validation data BraTS (V), and over the clinically acquired MRI scans of OPG patients gathered into two datasets, containing 22 and 51 patients with known ground-truth segmentations.
- We experimentally verify the possibility of employing the deep learning models trained over the OPG MRIs to segment LGG/HGG

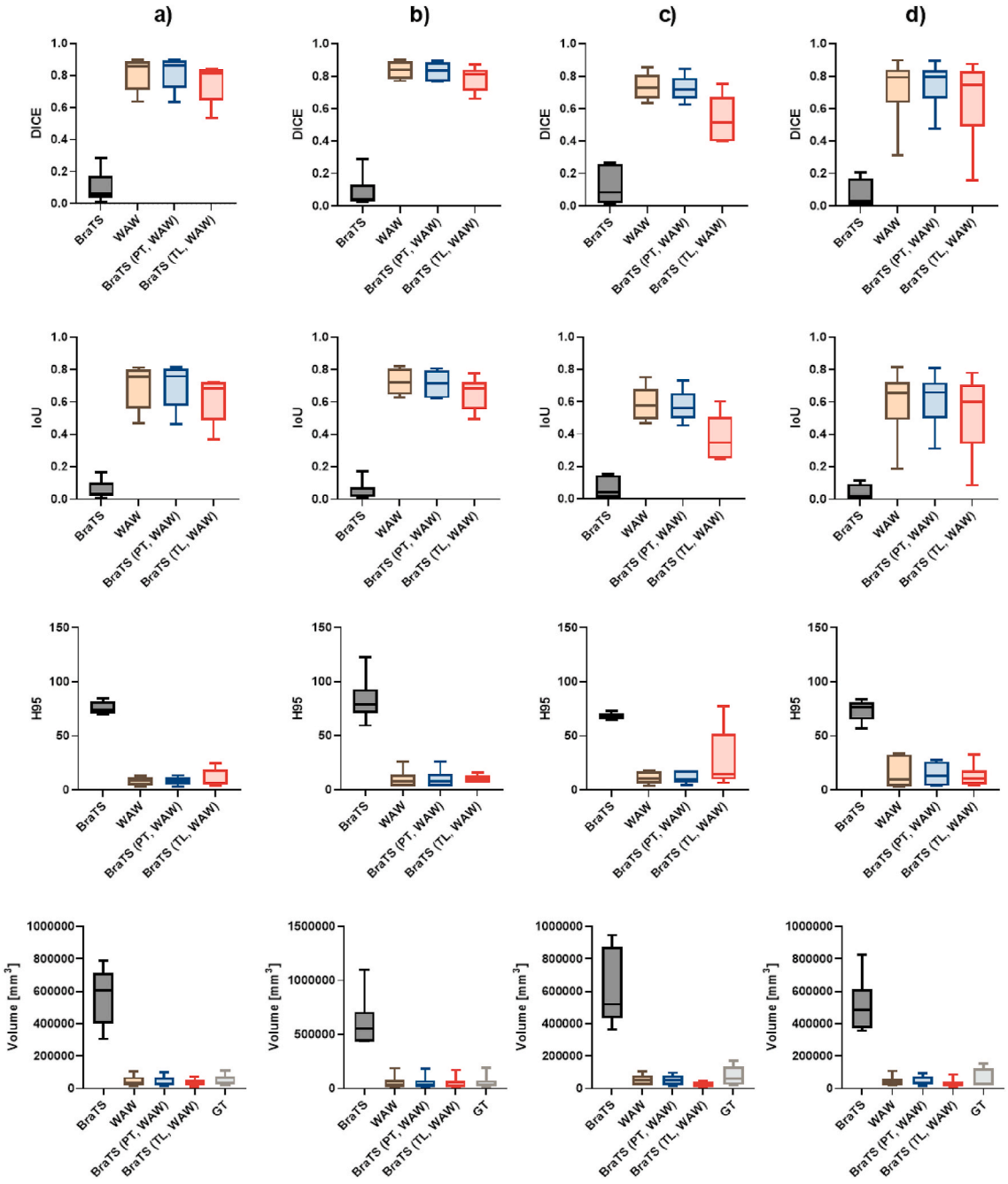

Fig. 4. Segmentation results obtained using the nnU-Net models trained in different settings over all test WAW folds (a–d: folds 1–4), quantified by DICE, IoU, and H95 (top three rows). We present the WT volume (in mm^3) obtained using all investigated models and compare it with the ground-truth (GT) WT volume calculated for manual tumor delineations (last row).

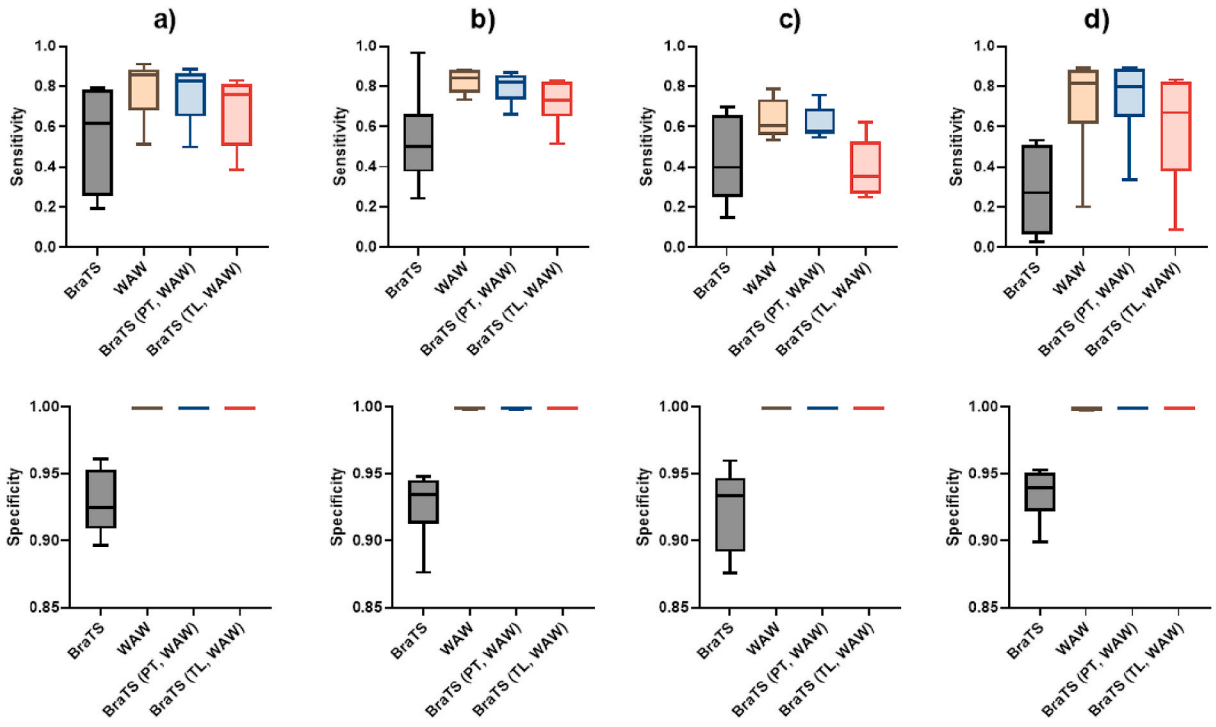

Fig. 5. Segmentation results obtained using the nnU-Net models trained in different settings over all test WAW folds (a–d: folds 1–4), quantified by sensitivity (first row) and specificity (second row).

MRIs and vice versa. To the best of our knowledge, this is the first study in which the cross-task (OPG vs. LGG/HGG segmentation) generalization capabilities of such algorithms have been quantitatively and qualitatively investigated.

- We make the architectures of our models open-sourced in order to ensure full reproducibility of the suggested method over other MRI data.

2. Materials and methods

2.1. Patient cohorts

We collected four cohorts of patients, two pre-operative GBM cohorts (adults), and two OPG cohorts (children)—the descriptive statistics of these datasets are gathered in Table 2. The BraTS pre-operative cohorts, referred to as BraTS 2020 (Tr) and BraTS 2020 (V) with 369 and 125 MRIs, respectively, include routine clinically acquired pre-operative mpMRI scans of GBM/high-grade gliomas and low-grade gliomas with pathologically confirmed diagnosis, captured in 19 institutions (years 2012–2020). In BraTS 2020 (Tr), we have 293 HGG and 76 LGG MRIs, whereas the number of HGG/LGG patients is unknown for BraTS 2020 (V). The MRIs include the native T1 (originally acquired in sagittal or axial orientation), T1c (Gadolinium, 3D axial), T2 (2D axial) and FLAIR (2D axial, coronal or sagittal) sequences (all with variable slice thickness). The scans were acquired using clinical conditions followed in each institution, hence different equipment and imaging protocols were exploited. Therefore, the BraTS cohorts reflect very heterogeneous MRIs of varying quality. All sequences were co-registered to a common anatomical template [34], and further resampled to 1 mm^3 and skull-stripped. The data is publicly available through the Image Processing Portal of the Center for Biomedical Image Computing and Analytics (CBICA) at the University of Pennsylvania, USA (https://ipp.cbica.upenn.edu/; last accessed: September 9, 2021).

The BraTS cohorts were segmented by one to four readers who followed the same procedure (delivered to each contributing institution), and all annotations were reviewed by experienced neuro-radiologists. The delineated tumor sub-regions included the active (enhancing) tumor (ET), showing hyperintensity regions in post-contrast T1 when compared to the native T1, the gross tumor, also referred to as the tumor core (TC—the bulk of tumor that is typically resected, including ET, alongside the necrotic and non-enhancing parts of the tumor, being the hypo-intense regions in post-contrast T1 when compared to the native T1), and the complete (whole) tumor extent (WT), entailing TC and the peritumoral edematous/invaded tissue (ED) that is typically manifested as hyperintense signal in FLAIR (excluding the contralateral and periventricular regions, unless they are contiguous with peritumoral edema, as they commonly represent age-associated demyelination or chronic microvascular changes rather than tumor infiltration [35]). The final ground-truth segmentations were reviewed for consistency by a single board-certified neuro-radiologist (more than 15 years of experience, YOE) [12]. The BraTS organizers make the ground-truth delineations available only for the training MRIs—therefore, only BraTS 2020 (Tr) can be used for training models, whereas BraTS 2020 (V) is treated as an independent validation set over which the generalization abilities of the algorithms are quantified (the quality metrics are calculated by the BraTS evaluation server at https://ipp.cbica.upenn.edu/, as the ground-truth masks are not publicly available; last access: October 4, 2021).

The OPG cohorts, referred to as WAW and WAW_{Test}, as these datasets were acquired in Warsaw, Poland, include 22 patients (13 girls and 9 boys, average age: 7.5 years, median age: 8 years) and 51 patients (29 girls and 22 boys, average age: 9 years, median age: 8 years), respectively. For these patients, the T2 and FLAIR MRI sequences captured axially using a 1.5 T scanner at the Children's Memorial Health Institute, Warsaw, Poland (years 2016–2021). The MRIs were segmented by one reader (15 YOE), and the manual delineations were reviewed by the expert radiologist (35 YOE). Gliomas of the visual pathway show morphological heterogeneity due to the presence of the solid and cystic parts. In T2 and FLAIR images, significant differences in signals are visible for both parts, with the solid and cystic areas becoming hyperintense in T2. On the other hand, the solid part is commonly manifested as iso- or hyperintense areas in FLAIR, whereas the cystic part can

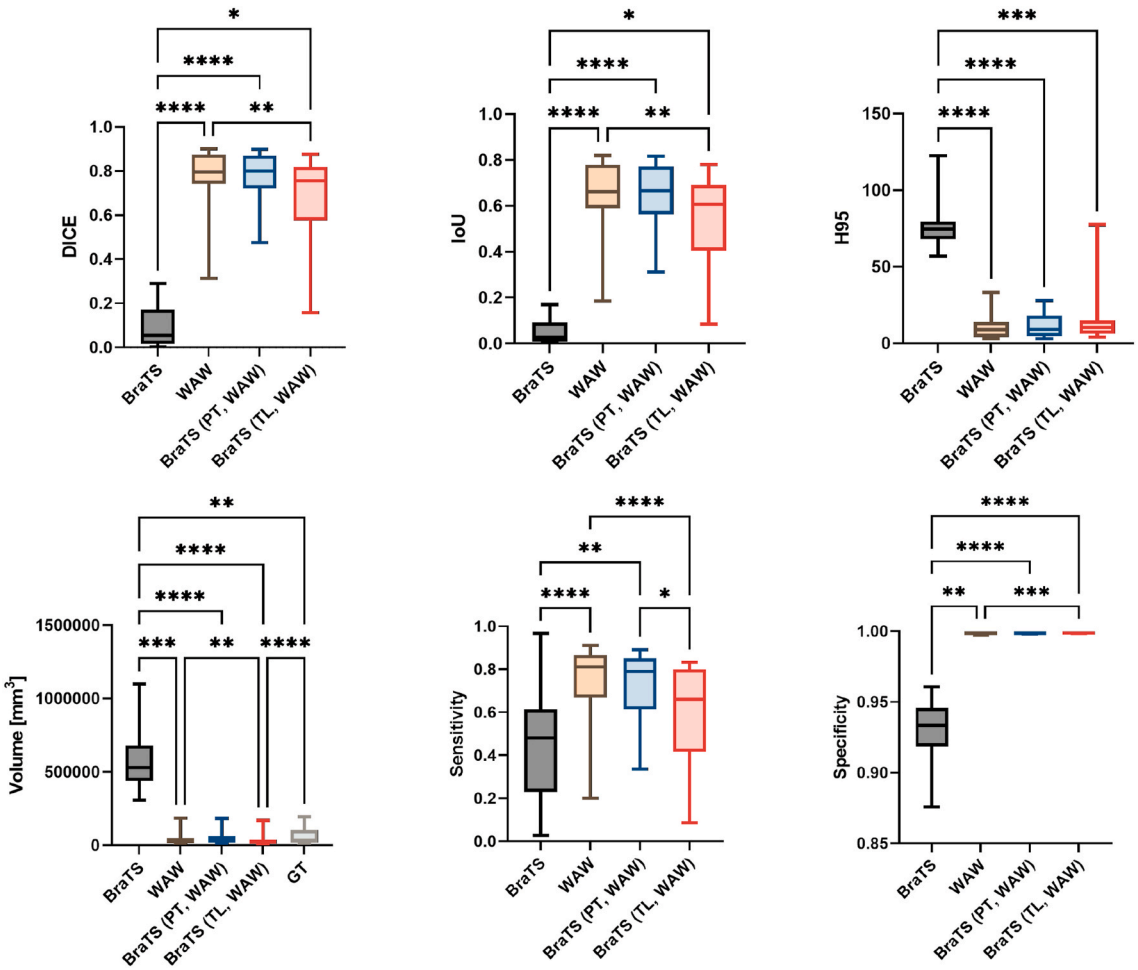

Fig. 6. Segmentation results obtained using the nnU-Net models trained in different settings over all test WAW patients, quantified by DICE, IoU, H95 (in mm), sensitivity, and specificity. We also present the WT volume (in mm^3) obtained using all investigated models and compare it with the ground-truth (GT) WT volume calculated for manual tumor delineations. The results of the statistical tests (Friedman's test with post-hoc Dunn's) aimed at verifying if the differences across models are statistically significant are presented as: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.000\ 1$). If the differences across models are not statistically significant (at $p < 0.05$), no asterisks are rendered.

become hypo-, iso- or hyperintense in FLAIR. The manual segmentation process involved the analysis of both sequences in order to accurately delineate the boundaries of the neoplastic lesions (T1-weighted images after the contrast agent administration were not considered because of the visible enhancement in the solid part only, which could negatively affect the visual analysis of the tumor's boundary).

2.1.1. Training and test MRIs

All MRIs were divided into training and test scans, with the latter used for quantifying the generalization abilities of the algorithm over the unseen data. Here, the entire BraTS 2020 (Tr) cohort is used as a training sample, whereas BraTS 2020 (V) is utilized for assessing the segmentation quality of models over the LGG/HGG patients—although we focus on segmenting OPG in children in this study, we also aim at understanding the segmentation abilities of deep learning algorithms targeted at OPG for fundamentally different brain lesions. This could ultimately help us design techniques that are applicable in different clinical settings. Due to a relatively small size of the WAW dataset, we follow the four-fold cross-validation approach in the case of this cohort. The WAW dataset is split into four non-overlapping folds (F1–F4) at the patient level with stratification reflecting the distribution of the whole-tumor volume in WAW (Fig. 1), and each fold is treated as the unseen test set exactly once. Finally, all WAW_{Test} MRIs are used as the unseen test set to quantify the generalization abilities of the deep models (these MRIs are never utilized during the training process). Although the BraTS 2020 dataset includes more modalities, we focus on T2 and FLAIR sequences in this study, as they are available for all cohorts.

In Fig. 2, we concisely present the training and test MRIs that are exploited in this study. For the WAW dataset, the number of patients (MRIs) in the training/test subsamples vary according to the specific fold. Finally, each WAW patient is utilized exactly once as a test MRI, therefore it is included in a single test fold (each fold F1–F4 becomes a test fold once during the experimentation). Overall, the segmentation performance of the models is assessed over OPG, LGG and HGG MRIs, captured at multiple institutions, following different acquisition procedures. This approach enables us to quantify the generalization abilities of the underlying models in various clinical settings.

2.2. Deep learning for segmenting OPG from MRI

Our segmentation engine is built upon an extremely successful nnU-Net [23] which established the state of the art in a range of medical segmentation tasks in various modalities and organs, including, among others, brain tumors, liver, lung, prostate, kidney, or spleen—nnU-Net was shown to be outperforming most specialized algorithms, as presented in Fig. 3 in Ref. [23]. The nnU-Net algorithm is a deep learning segmentation method that automatically adapts itself based on the characteristics of the training data and target segmentation problem.

Table 3
Segmentation results obtained using the nnU-Net models trained in different settings over all test WAW patients, quantified by DICE, IoU, H95, sensitivity, and specificity. The best results for each aggregation: 25% and 75% percentiles, median, mean and lower and upper 95% confidence interval (CI) of mean, are bold, and the second best results are underlined.

Quality metric		BraTS	WAW	BraTS (PT, WAW)	BraTS (TL, WAW)
DICE	25% Percentile	0.016	**0.741**	0.721	0.575
	Median	0.055	0.796	**0.799**	0.754
	75% Percentile	0.170	**0.876**	0.871	0.818
	Mean	0.093	0.778	**0.781**	0.686
	Lower 95% CI of mean	0.049	0.720	**0.734**	0.603
	Upper 95% CI of mean	0.137	**0.836**	0.829	0.770
IoU	25% Percentile	0.008	**0.589**	0.563	0.403
	Median	0.028	0.661	**0.666**	0.605
	75% Percentile	0.093	**0.779**	0.771	0.692
	Mean	0.052	0.652	**0.652**	0.549
	Lower 95% CI of mean	0.026	0.585	**0.593**	0.463
	Upper 95% CI of mean	0.077	**0.719**	0.712	0.635
H95	25% Percentile	67.96	**3.89**	4.58	6.11
	Median	74.40	**8.98**	9.02	10.33
	75% Percentile	79.54	**13.96**	18.09	14.83
	Mean	75.38	11.22	**11.10**	14.80
	Lower 95% CI of mean	69.66	**7.25**	7.59	7.77
	Upper 95% CI of mean	81.10	15.20	**14.61**	21.83
Sensitivity	25% Percentile	0.2305	**0.6679**	0.6148	0.4166
	Median	0.4801	**0.8109**	0.7892	0.6599
	75% Percentile	0.6131	**0.8657**	0.8503	0.7994
	Mean	0.4442	**0.7505**	0.7355	0.5997
	Lower 95% CI of mean	0.3330	**0.6744**	0.6691	0.5019
	Upper 95% CI of mean	0.5555	**0.8266**	0.8018	0.6975
Specificity	25% Percentile	0.9185	0.9987	0.9990	**0.9995**
	Median	0.9336	0.9995	0.9996	**0.9997**
	75% Percentile	0.9458	0.9996	0.9997	**0.9998**
	Mean	0.9287	0.9992	0.9993	**0.9996**
	Lower 95% CI of mean	0.9179	0.9989	0.9991	**0.9994**
	Upper 95% CI of mean	0.9394	0.9995	0.9996	**0.9998**

This configuration includes data pre-processing, basic post-processing, network architecture and training parameters[1]—the detailed hyper-parameters of the elaborated nnU-Net model are available at https://gitlab.com/jnalepa/OPG. Ultimately, the generated network is a 2D U-Net consisting of an encoder and a decoder (forming the contractive and expanding paths) which are interconnected by skip connections. This U-shape architecture allows us to propagate high-level features extracted in the contractive path through the higher-resolution layers in the expanding path, effectively performing multi-scale analysis of the input scan. The algorithm operates on co-registered T2 and FLAIR sequences, and outputs the mask with the voxels annotated as either healthy or abnormal (tumorous). Therefore, we focus on the binary whole-tumor classification. Since the nnU-Net automatically configures the pre-processing routines, we do not employ any additional pre-processing steps.

2.2.1. *Training strategies: pre-training and transfer learning*

As there are no representative datasets in the literature that could be used for training large capacity deep learners, we introduce two training strategies for the OPG segmentation in this study—in both ones, we effectively execute two separate training processes sequentially (Fig. 3). In the *pre-training strategy* (PT), we exploit the available ground-truth dataset of a larger size and of similar characteristics to the OPG MRIs to avoid a significant domain shift (i.e., we utilize the LGG/HGG MRIs with manual delineations), referred to as the *source data*, to pre-train nnU-Nets.[2] Here, the source data is of the same modality (T2 and FLAIR MRI sequences) as the *target data*, but reflects a different segmentation task (LGG/HGG segmentation). Also, the amount of the source data should be substantially larger than the target data, as it is utilized for pre-training the segmentation engine, therefore for elaborating high-quality deep feature extractors. The pre-trained nnU-Net model serves as a starting point for the re-training step, in which (much smaller) target data is used for re-training the network to tackle the task of interest (OPG segmentation). In this strategy, the entire pre-trained nnU-Net (meaning all its trainable weights) is adapted in the training process which is finally performed over the OPG training MRIs.

In the *transfer learning strategy* (TL), the nnU-Net is pre-trained over the source data (similarly to the pre-training strategy). Afterwards, the classification part of the architecture is fine-tuned over the target OPG training data. Therefore, only the parameters of the final (classification) layers are fine-tuned, whereas all other (feature extraction) layers are kept unchanged during the second training phase. In both strategies, the hyper-parameters of the training process, including the optimizer, maximum number of epochs, and the data augmentation routines are fixed for both sequential nnU-Net training phases. Therefore, the main difference between PT and TL is the "depth" of parameters' update in the pre-trained nnU-Net. In TL, we fine-tune only a relatively small subset of 4.15 million parameters, out of 18.67 million trainable weights (therefore, only 22.24% parameters are updated),[3] whereas in PT all parameters (18.67 million) are updated. Finally, the models that are obtained through employing both strategies are confronted with those trained solely on the source or target data, being BraTS 2020 (Tr) and WAW (training folds), respectively.

3. Experimental results

Our method was implemented in Python 3.7 with the TensorFlow 2.3 backend. The R package IRR (Inter Rater Reliability, version 0.84.1) was used for calculating the Intraclass Correlation Coefficient (ICC), whereas GraphPad Prism 9.2.0 was utilized for other statistical analysis. The segmentation models are built upon a well-established open-sourced nnU-Net framework publicly available at https://github.com/MIC-DKFZ/nnUNet. To ensure full reproducibility of the algorithm, the architectural diagrams of our final nnU-Net models, together with the hyper-parameters of pre-processing routines elaborated by the nnU-Net engine and all training curves are accessible at https://gitlab.com/jnalepa/OPG. The experiments were executed on a personal computer equipped with an NVIDIA TITAN X (12 GB) and Intel i7-6850K CPU (3.60 GHz) CPU.

The nnU-Net is trained using stochastic gradient descent with

[1] To fully understand all pivotal aspects of the nnU-Net framework, we refer to an excellent Fig. 2 presented in Ref. [23], in which Isensee et al. clearly discussed which algorithm's properties are extracted from the incoming dataset for a given medical image segmentation task, how they are used to infer specific parameters of the pipeline, which parameters do not require such adaptation, and how the ensemble is ultimately built.

[2] Note that the training strategies suggested in this work are model-agnostic, and can be exploited to train any underlying deep architecture. Such approaches have indeed been utilized in other medical image analysis tasks [36], and also in other image analysis domains [37].

[3] See the layers after the concatenation of features from the contracting and expanding paths in the nnU-Net architecture deposited at https://gitlab.com/jnalepa/OPG.

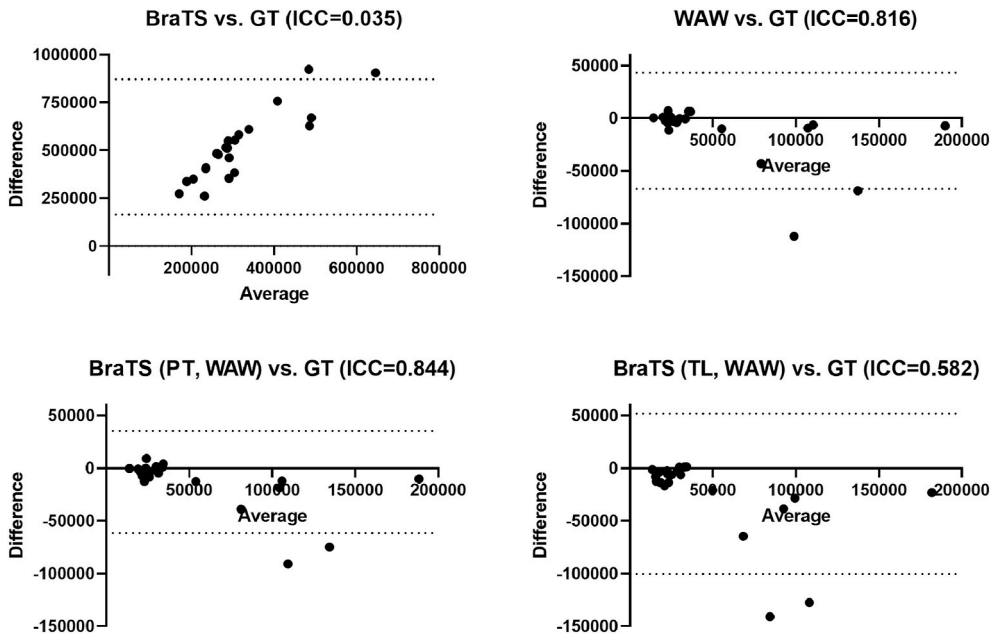

Fig. 7. Bland-Altman plots for the nnU-Net models trained in different settings over all test WAW patients showing the agreement between the automatically extracted and ground-truth WT volume, quantified using the Intraclass Correlation Coefficient (ICC).

Nesterov momentum ($\mu = 0.99$) for 1000 epochs. The batch included the 320×320 patches (bath size of 32) or 192×160 (batch size of 108) for the nnU-Net model trained over WAW and BraTS 2020 (Tr) (contatining 29.97 and 18.67 million trainable parameters, respectively). Here, the latter model is fine-tuned in either PT or TL strategy, and 250 batches were processed within each epoch. Training-time data augmentation encompassed random patch scaling within (0.7, 1.4), random rotation, random gamma correction within (0.7, 1.5), and random mirroring [38]. The models were trained with the loss function being the averaged cross-entropy and soft DICE, where both are commonly used in semantic segmentation [39].

Overall, we investigate 13 nnU-Net models trained (or fine-tuned) over different training samples:

- **BraTS**—the nnU-Net model trained over the BraTS 2020 (Tr) cohort. This training strategy leads to obtaining **one nnU-Net.** Training this model took 71 h.
- **WAW**—the nnU-Net model trained over the training folds extracted from the WAW cohort. This training strategy leads to obtaining **four nnU-Nets** trained over different WAW training samples. Training these model took 54 h (per model).
- **BraTS (PT, WAW)**—the nnU-Net model trained in the PT strategy. It is pre-trained over the BraTS 2020 (Tr) cohort, and later re-trained using the training folds extracted from the WAW cohort (see Section 2.2.1 for more details). This training strategy leads to obtaining **four nnU-Nets** re-trained over different WAW training samples. Training this model took from 45 up to 72 h (mean: 60 h), depending on the fold.
- **BraTS (TL, WAW)**—the nnU-Net model trained in the TL strategy. Here, it is pre-trained over the BraTS 2020 (Tr) cohort, and later fine-tuned using the training folds extracted from the WAW cohort (Section 2.2.1). This training strategy leads to obtaining **four nnU-Nets** fine-tuned over different WAW training samples. Training this model took from 39 up to 45 h (mean: 41.5 h), depending on the fold.

All of the above-discussed algorithms were capable of delivering very fast operation over the unseen WAW and BraTS (V) MRIs, and the average end-to-end analysis time (of a single MRI) amounted to less than 10 s for each model.

To evaluate the segmentation, we exploited the DICE coefficient, Jaccard's index (also referred to as the Intersection over Union, IoU), sensitivity, specificity (the metrics range from 0 to 1, and the larger those measures become, the better performance is obtained, with 1 denoting the perfect score), alongside the 95th percentile of the Hausdorff distance (H95; the smaller, the better)—the 95th percentile of this metric is commonly used to prune the outliers. The DICE score is:

$$\mathrm{DICE}(\mathrm{P},\mathrm{GT}) = \frac{2\cdot|P \cap GT|}{|P| + |GT|} = \frac{2\cdot\mathrm{TP}}{2\cdot\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{1}$$

whereas IoU becomes:

$$\mathrm{IoU}(\mathrm{P},\mathrm{GT}) = \frac{|P \cap GT|}{|P \cup GT|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{2}$$

where P and GT are two segmentations (predicted and ground truth), and TP, FP, and FN are the numbers of true positives, false positives, and false negatives. Although both DICE and IoU are the overlap metrics, we can observe that IoU will penalize single instances of wrong segmentation more than DICE, hence IoU tends to quantify the "worst" case average performance. Sensitivity is the percentage of lesion pixels correctly classified as lesions $\left(\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{TP}}\right)$, specificity becomes the percentage of correctly classified healthy voxels out of all healthy voxels ($\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$; TN denotes true negatives). Overall, we extract a set of quality metrics that enable us to thoroughly evaluate the performance of the deep models, both in the context of overlap of the manual and automatic segmentation masks (DICE, IoU), and the contour similarities (H95). It is important to emphasize that high-quality contours are pivotal in clinical settings, where uni- or bidimensional measurements of the tumor areas are extracted to quantify the tumor progression and patient's response [11], as the quality of such measurements can be easily affected by wrong contouring.

3.1. Experiment 1: Segmentation of OPG

In this experiment, we focus on quantifying the OPG segmentation abilities of the nnU-Net models trained in different settings (Section 2.2.1), and over various ground-truth MRI datasets. For the WAW

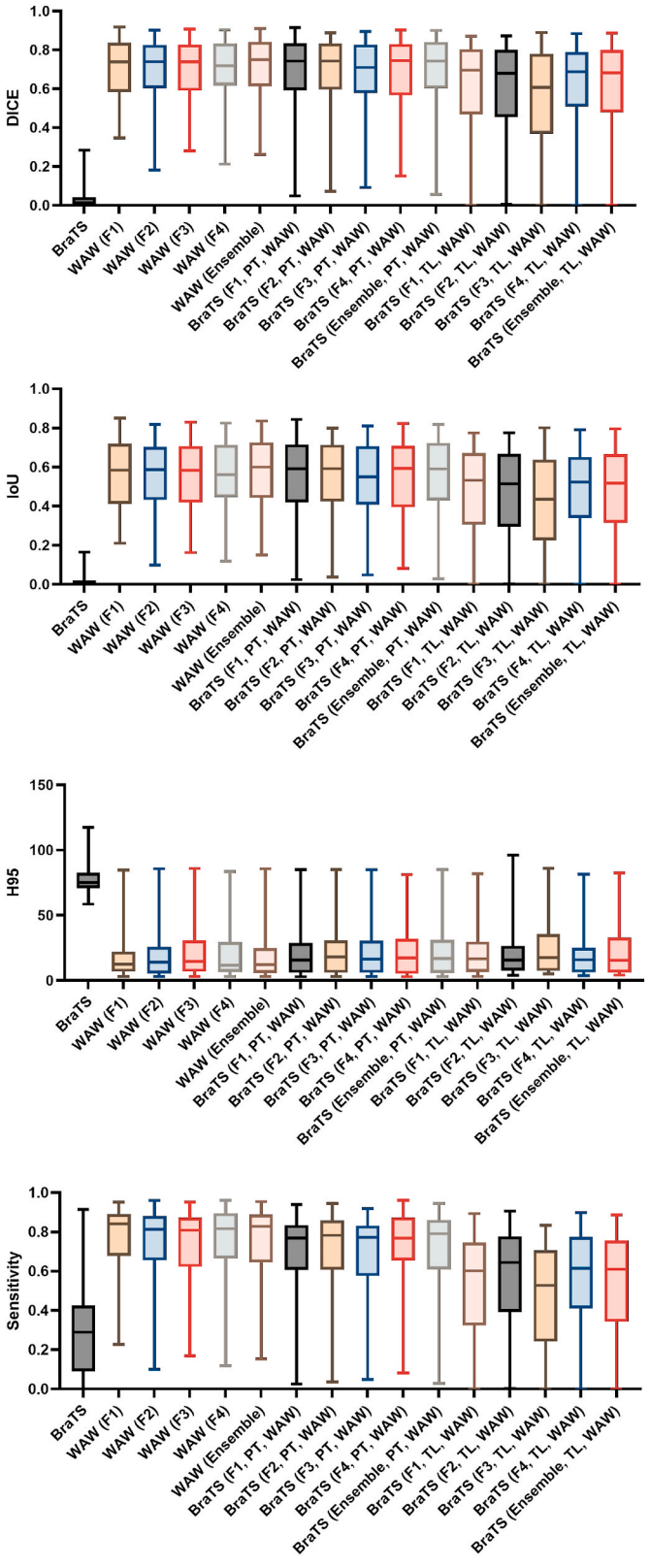

Fig. 8. Segmentation results obtained using the nnU-Net models trained in different settings over all WAW_{Test} patients, quantified by DICE, IoU, H95, and sensitivity.

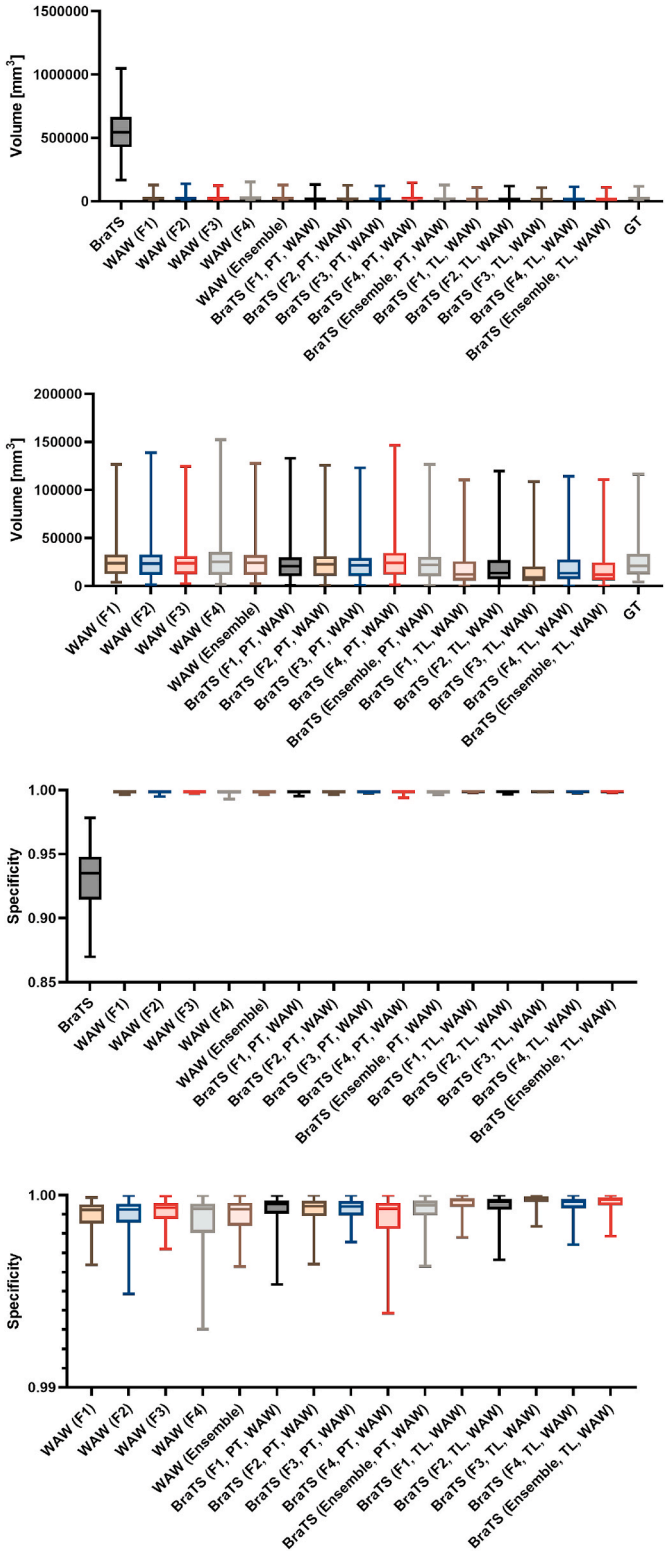

Fig. 9. Segmentation results obtained using the nnU-Net models trained in different settings over all WAW_{Test} patients, quantified by the whole-tumor volume and specificity. For both metrics, we present the plots with and without the BraTS model to ensure readability.

Table 4
Segmentation results obtained using the nnU-Net ensembles trained in different settings and the model trained over the BraTS training data over all test WAW_{Test} patients, quantified by DICE, IoU, H95, sensitivity, and specificity. The best results for each aggregation: 25% and 75% percentiles, median, mean and lower and upper 95% confidence interval (CI) of mean, are bold, and the second best results are underlined.

Quality metric		BraTS	WAW	BraTS (PT, WAW)	BraTS (TL, WAW)
DICE	25% Percentile	0.006	**0.613**	0.601	0.478
	Median	0.014	**0.750**	0.742	0.682
	75% Percentile	0.041	**0.841**	0.840	0.799
	Mean	0.036	**0.713**	0.694	0.614
	Lower 95% CI of mean	0.020	**0.670**	0.643	0.549
	Upper 95% CI of mean	0.052	**0.756**	0.745	0.680
IoU	25% Percentile	0.003	**0.442**	0.430	0.314
	Median	0.007	**0.599**	0.590	0.517
	75% Percentile	0.021	**0.725**	0.723	0.665
	Mean	0.019	**0.574**	0.557	0.479
	Lower 95% CI of mean	0.010	**0.525**	0.503	0.417
	Upper 95% CI of mean	0.028	**0.622**	0.612	0.541
H95	25% Percentile	70.65	5.75	**5.70**	6.02
	Median	75.05	**12.20**	16.98	15.58
	75% Percentile	82.72	**24.77**	31.11	32.92
	Mean	77.20	**17.04**	20.25	20.50
	Lower 95% CI of mean	74.01	**12.66**	15.53	15.48
	Upper 95% CI of mean	80.40	**21.41**	24.97	25.52
Sensitivity	25% Percentile	0.0908	**0.6447**	0.6098	0.3430
	Median	0.2896	**0.8291**	0.7908	0.6107
	75% Percentile	0.4257	**0.8876**	0.8617	0.7562
	Mean	0.2860	**0.7627**	0.7089	0.5414
	Lower 95% CI of mean	0.2207	**0.7145**	0.6509	0.4715
	Upper 95% CI of mean	0.3514	**0.8110**	0.7668	0.6112
Specificity	25% Percentile	0.9145	0.9984	0.9990	**0.9995**
	Median	0.9351	0.9993	0.9995	**0.9998**
	75% Percentile	0.9479	0.9996	0.9997	**0.9999**
	Mean	0.9325	0.9990	0.9992	**0.9996**
	Lower 95% CI of mean	0.9265	0.9988	0.9990	**0.9995**
	Upper 95% CI of mean	0.9386	0.9993	0.9994	**0.9997**

patients, the models trained over the BraTS training data failed to deliver accurate WT segmentation in all test folds (Figs. 4 and 5). The segmentation results aggregated across all WAW patients[4] are presented in Fig. 6 and Table 3. The results for the nnU-Nets trained over the WAW training patients (for each fold, as presented in Fig. 4, and collectively for all test folds in Fig. 6 and Table 3) indicate a significant increase in the segmentation abilities of the models. Similarly, training in the PT and TL strategies lead to notable improvements in all measures (the mean DICE was improved by ↑0.688 and ↑0.593 for PT and TL, respectively—analogously, we can observe enhancements of the IoU scores; the mean H95 dropped by ↓64.28 mm and ↓60.58 mm, which shows that the automated contouring became much more similar to the manually-delineated whole-tumor regions) when compared to the nnU-Nets trained solely on the BraTS training set. Other metrics based on the analysis of the confusion matrix highlight very low specificity and sensitivity of the BraTS nnU-Net model, manifesting its inability to accurately delineate tumors from OPG image data.

The Friedman's test, followed by the post-hoc Dunn's shows that the differences between the networks trained solely on BraTS (Tr) and all deep models for which the training process involved the OPG training data, i.e., the models learned from the WAW training data only, or utilizing the PT and TL strategies, are statistically significant in virtually all cases—the only exception is the difference in sensitivity of the BraTS and BraTS (TL, WAW) nnU-Nets, for which $p > 0.05$ (Fig. 6). The differences in quality metrics are not statistically significant for the BraTS (TL, WAW) and BraTS (PT, WAW) deep neural network models at $p < 0.05$ (excluding sensitivity).

The automatic volumetric measurements obtained for all WAW test patients (in mm^3) delivered by the nnU-Net model trained over the BraTS training MRIs were in strong disagreement with the ground truth segmentations (ICC: 0.035), with the Bland-Altman plot indicating significant over-segmentation of the OPGs in these patients (Fig. 7). The ICC values were notably improved for all other nnU-Nets, and amounted to 0.582, 0.816 and 0.844 ($p < 0.001$) for the models trained in the TL strategy, over the WAW training MRIs only, and in the PT strategy, respectively, indicating strong up to almost perfect agreement between the manual and automatic volumetric measurements. The most significant discrepancies in the ground-truth and automatically calculated whole-tumor volumes, using the nnU-Nets that benefit from the available WAW training MRIs, were observed for the largest lesions, as presented in the corresponding Bland-Altman plots rendered in Fig. 7.

To verify the importance of specific MRI sequences on the operational abilities of the nnU-Net models, we voluntarily removed one sequence (either T2 or FLAIR) from the test MRIs. Overall, the quality of segmentation was catastrophically deteriorated in both cases. The mean DICE dropped from 0.093 to 0.007 for the BraTS model, from 0.778 to 0.152 for WAW, from 0.781 to 0.000 for BraTS (PT, WAW), and from 0.686 to 0.097 for BraTS (TL, WAW) once FLAIR was removed. In contrary, the same metric values decreased to 0.007, 0.004, 0.000, and 0.000 for BraTS, WAW, BraTS (PT, WAW) and BraTS (TL, WAW), respectively, after rejecting the T2 sequence.

The generalization capabilities of all deep models are quantified over the unseen WAW_{Test} MRIs that were never used during the training process. The results in Figs. 8 and 9 confirm the previous observations that that the BraTS model delivers the worst performance over the OPG patients, as shown by all quality metrics. The aggregated results (obtained by the BraTS model and all investigated ensembles) gathered in Table 4 manifest significant improvements of the proposed segmentation engines when compared to the one trained over the BraTS training MRIs only (all improvements are statistically significant).

The Bland-Altman plots obtained for the WAW_{Test} MRIs indicate massively large over-estimation of the whole-tumor volume delivered by the BraTS model (Fig. 10). On the other hand, all ensembles trained in the suggested strategies were in almost perfect agreement with the human readers, resulting in ICC up to 0.958 for WAW (Ensemble). The plots do not present any clear tendency to under/over-estimate the whole-tumor volume by WAW (Ensemble) and BraTS (Ensemble, PT, WAW). For BraTS (Ensemble, TL, WAW), we can observe that this ensemble tends to elaborate lower volumes (hence under-segments the tumors) than those captured in the ground-truth data, but still the volumes are in strong agreement with the raters (ICC amounted to 0.873).

3.2. Experiment 2: Segmentation of LGG/HGG

In this experiment, we verify the LGG/HGG segmentation abilities of

[4] Note that we present the results obtained for the test OPG patients, meaning that each WAW patient was included in the test fold exactly once—it means that the patients belonging to the test fold 1 were segmented using nnU-Nets trained over the training patients from folds 2–4; the patients belonging to the test fold 2 were segmented using nnU-Nets trained over the training patients from folds 1, 3, and 4, and so forth. Therefore, in Fig. 6 and Table 3, we collectively present the quality metrics obtained for all test cases. To this end, we do not have any training-test information leak which could lead to over-optimistic segmentation results.

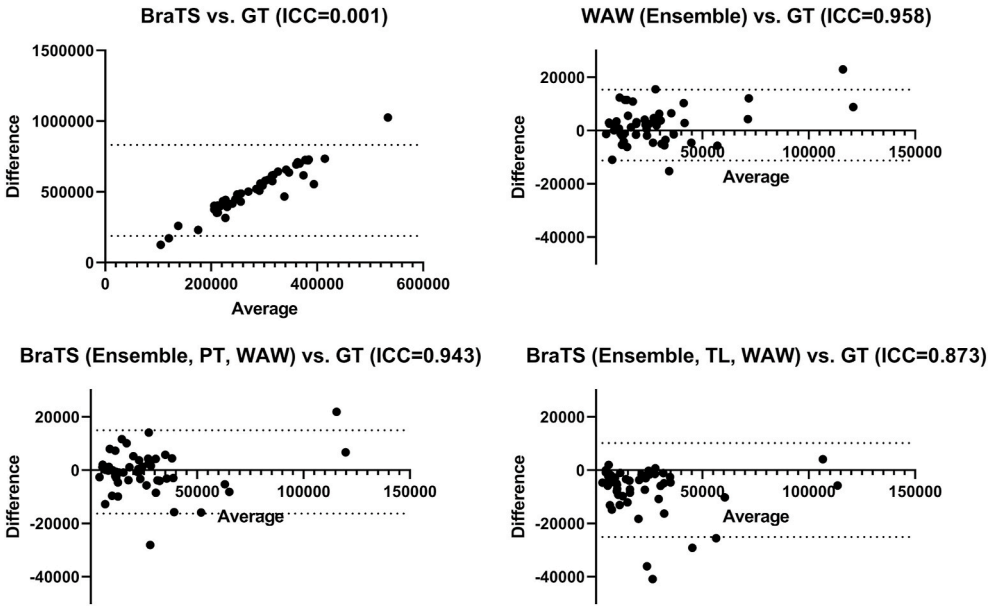

Fig. 10. Bland-Altman plots for the model trained over BraTS (Tr) and nnU-Net ensembles trained in different settings showing the agreement between the automatically extracted and ground-truth WT volume in all WAW_{Test} patients, quantified using the Intraclass Correlation Coefficient (ICC).

the nnU-Net models trained over the BraTS training data, together with those that were trained utilizing the WAW patients only, and the models trained in the PT and TL strategies. Since the generalization abilities are assessed over the BraTS validation MRIs (125 MRIs corresponding to 125 pre-surgery patients), the entire WAW data could be used for training (here, we exploit the T2 and FLAIR sequences only, in order to follow the same protocol as for OPGs). We elaborated the models for all folds separately (referred to as the F1–F4 models), and then built a segmentation ensemble which encompasses all base models trained in a specific way (over WAW only, or in the PT and TL strategies). The base deep models are assembled in the ensemble segmentation engines by averaging softmax probabilities.

The nnU-Net model trained over BraTS delivered the best-quality segmentation of unseen LGG/HGG validation patients, with the mean DICE of 0.896 (95% CI: 0.882–0.910, median: 0.921), and the mean H95 of 5.56 mm (95% CI: 3.98–7.14 mm, median: 3.61 mm). On the other hand, the nnU-Net trained over the third WAW fold (F3) obtained the lowest mean DICE: 0.414 (95% CI: 0.370–0.459, median: 0.430), and the corresponding mean H95 of 41.28 mm (95% CI: 37.25–45.31 mm, median: 36.28 mm). The models trained in both PT and TL strategies elaborated better delineations of the LGG/HGG patients when compared to the WAW models (trained on the OPG MRIs only)—in PT, the model that utilized the third WAW fold (F3) for fine-tuning the pre-trained model obtained the mean DICE of 0.504 (95% CI: 0.451–0.556, median: 0.592; therefore, the mean DICE was improved by ↑0.09 comparing to the WAW nnU-Net model trained over F3), whereas in TL, the mean DICE of the corresponding nnU-Net amounted to 0.437 (95% CI: 0.389–0.485, median: 0.498; the mean DICE was improved by ↑0.023). An improvement can be observed in H95 for the PT strategy, where the mean value of this metric was decreased by ↓11.89 mm, whereas in TL, the mean H95 was increased by ↑7.07 mm. Building the ensemble segmentation engines helps significantly enhance the quality of the worst base models through benefiting from the better models in all training strategies, and delivered the mean DICE of 0.555 (95% CI: 0.507–0.603, median: 0.649), 0.563 (95% CI: 0.509–0.616, median: 0.701), 0.548 (95% CI: 0.499–0.598, median: 0.634) for the models trained over WAW, and in the PT and TL strategies. The corresponding improvements in mean DICE with respect to the worst base models amounted to ↑0.140, ↑0.059, and ↑0.111, respectively—all improvements are statistically significant (according to the Friedman's test with post-hoc Dunn's), with the p-values of $p < 0.000\ 1$, $p = 0.007\ 9$, and $p < 0.000\ 1$.

The perfect agreement between the automatically calculated and ground-truth WT volume obtained for the BraTS (Tr) training dataset (ICC: 1.000) was obtained using the nnU-Net model trained over the very same dataset, and the ensemble of nnU-Nets trained in the TL strategy (Fig. 12). Very high agreement was also observed for two other ensembles, trained over all WAW (OPG) MRIs, WAW (Ensemble), ICC: 0.622, and following the PT strategy, BraTS (Ensemble, PT, WAW), ICC: 0.681. It is worth noting that only the WAW ensemble did not utilize the BraTS (Tr) MRIs during its training, and these scans were kept aside and used as the independent test set for this algorithm—all other nnU-Net models are either trained entirely over BraTS (Tr), or utilize it to fine-tune their weights in PT and TL (therefore, the results may be treated as the agreement upper bound for these networks). After voluntarily removing one modality (T2 or FLAIR), the segmentation results rapidly dropped for all models, reaching the mean DICE close to zero.

4. Discussion

Automating the process of OPG detection and segmentation from MRI can impact the longitudinal radiological assessment of such scans to quantitatively estimate the tumor burden, and has potential to accelerate the diagnostic process. The irregularities of the OPG's boundaries and their varying locations, together with heterogeneous tumor voxels' intensities captured by MRI sequences lead to the lack of reproducibility of the manual delineation, and visible inter- and intra-rater disagreements (such disagreements vary due to the differences in relevant clinical experience of the raters). Therefore, designing and implementing accurate and reproducible segmentation methods is of high clinical importance as it can improve managing various types of lesions, including OPGs.

Our experimental study indicates that the LGG/HGG MRI scans are insufficient to train the nnU-Net models which can generalize well over OPG cases. Aggregating the segmentation results over all WAW and WAW_{Test} patients confirms this observation (Figs. 6 and Figs. 8–9, respectively), and shows that these models tend to significantly over-segment OPGs, leading to unacceptably large numbers of FPs that could not be used in clinical settings. In Fig. 13, an example FLAIR MRI frame is presented—for this test WAW patient, the model trained over

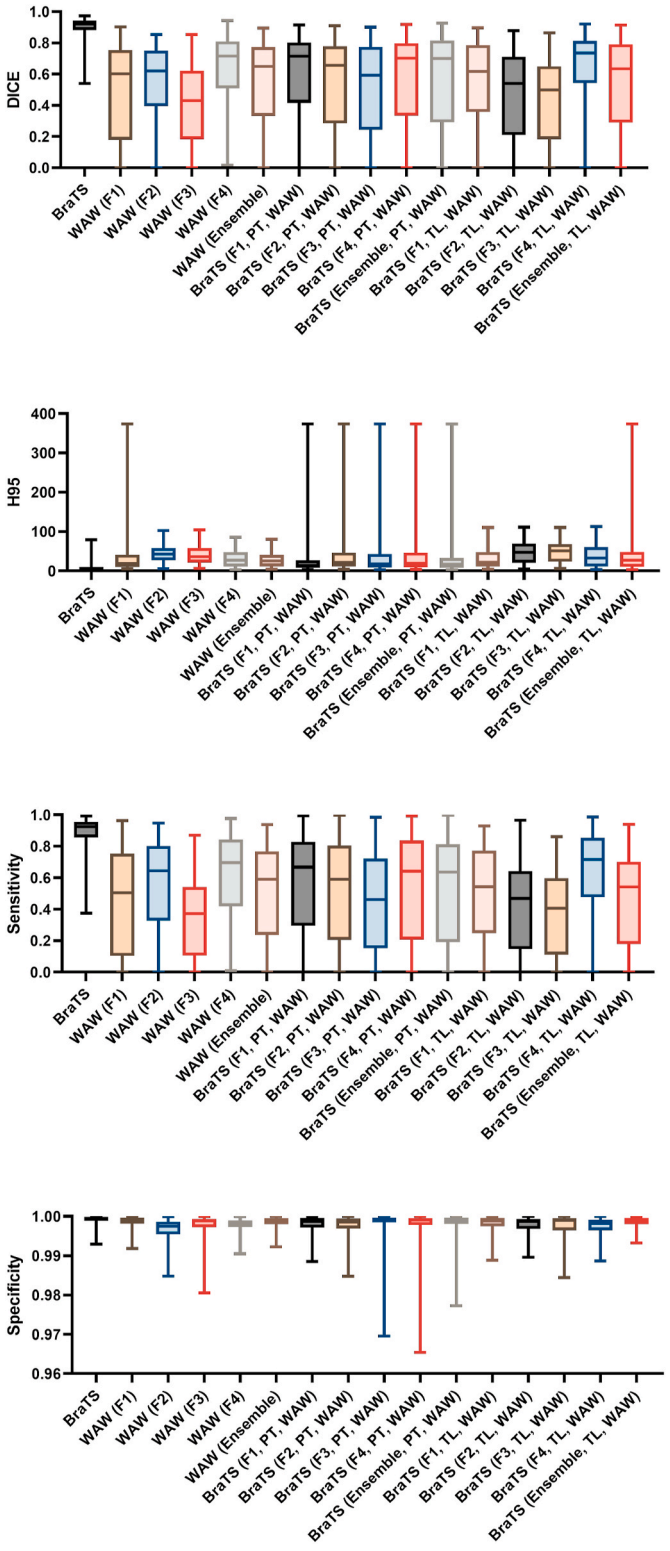

Fig. 11. Segmentation results obtained using the nnU-Net models trained in different settings over all validation BraTS patients, quantified by DICE, H95, sensitivity, and specificity, as returned by the validation server (other quality metrics are not calculated by it, hence we are unable to report them—the ground-truth segmentations are not publicly available).

the BraTS (Tr) MRIs obtained the DICE score of 0.290, while presenting significant over-segmentation of the tumor (note that a "large" value of DICE is related to the large area of this lesion). All other models, however, deliver much more precise segmentation, with the DICE score reaching 0.900 in this example. The same observation can be inferred from the segmented WAW_{Test} MRI rendered in Fig. 14, in which the BraTS model over-segmented the lesion as well (the DICE coefficient of 0.283 delivered by this model was the largest across all its DICE values obtained for the WAW_{Test} patients). Therefore, exploiting the available OPG MRI scans in the suggested training strategies, even if the number of such MRIs is extremely limited, helps build nnU-Nets that can effectively process unseen OPG MRIs (it is reflected in Table 3 for WAW, and in Table 4 for WAW_{Test} in all metrics; note that the "worst" case average performance quantified by IoU was also improved for the models utilizing the WAW training MRIs). Additional qualitative examples, gathered in Figs. 15 and 16 (for the WAW patients), and in Fig. 17 (for a selected WAW_{Test} patient) indicate the abilities of the models to appropriately locate OPGs of varying sizes and boundary characteristics. All of the investigated algorithms offered real-time inference, maintaining the analysis time of a single MRI below 10 s on average (for each algorithm, including the ensembles).

Although we did not confront the proposed algorithm with other state-of-the-art algorithms developed specifically for OPG segmentation (these techniques are gathered in Table 1), we emphasize that our engine is extensively exploiting the nnU-Net backend which was shown to be outperforming virtually all other deep learning architectures and classical machine learning and image analysis-based techniques in medical image segmentation tasks. We can therefore anticipate that employing our pipeline to the datasets summarized in Table 1 (using the training/test splits suggested in the corresponding papers) would very likely lead to comparable or better segmentation accuracy than reported in these papers. Although we are aware that comparing different algorithms over different test MRIs[5] may easily be misleading, we can observe that the mean DICE obtained using our nnU-Net trained in the PT strategy over the unseen test WAW MRIs amounted to 0.781, whereas the mean DICE for all WAW_{Test} patients was 0.713, 0.694, and 0.614 for the ensembles trained from the WAW MRIs exclusively (which are non-overlapping with WAW_{Test} MRIs), and in the PT and TL strategies, respectively. In comparison, the mean DICE ranged from 0.694 in Ref. [28], up to 0.761 in Ref. [25]. In Section 4, we indicate that building publicly available OPG sets containing freely accessible MRIs coupled with the ground truth is of utmost importance, as it would help the research community compare the algorithms in a thorough and fair way. Also, we hope that making our architectures open-sourced will encourage other groups to publish their architectures and/or implementations to ensure full reproducibility of the experiments.

To quantify the robustness of our approaches against missing MRI modalities, we voluntarily removed a single MRI sequence from the input WAW data (either T2 or FLAIR was removed, i.e., it was replaced by an empty image). It led to the observations which are in line with the previous works on automatic OPG segmentation—in practically all of the existing methods, both MRI sequences are exploited (Table 1). Making such algorithms robust against missing modalities is an interesting open issue, and addressing it could help increase the clinical utility of the software tool, especially if a T2/FLAIR sequence was not captured or is of low quality (for any reason). The same applies to the segmentation of other tumors, such as LGG/HGG lesions, as the algorithm was not robust against missing modalities in this case either.

Employing the nnU-Net models trained using a combination of the LGG/HGG and OPG MRIs showed that such deep learning algorithms are

[5] Unfortunately, the MRIs used for the OPG segmentation utilized in other works are not publicly available due to the privacy issues, thus we are unable to compare the performance of our techniques with the results reported in the corresponding works over the very same data.

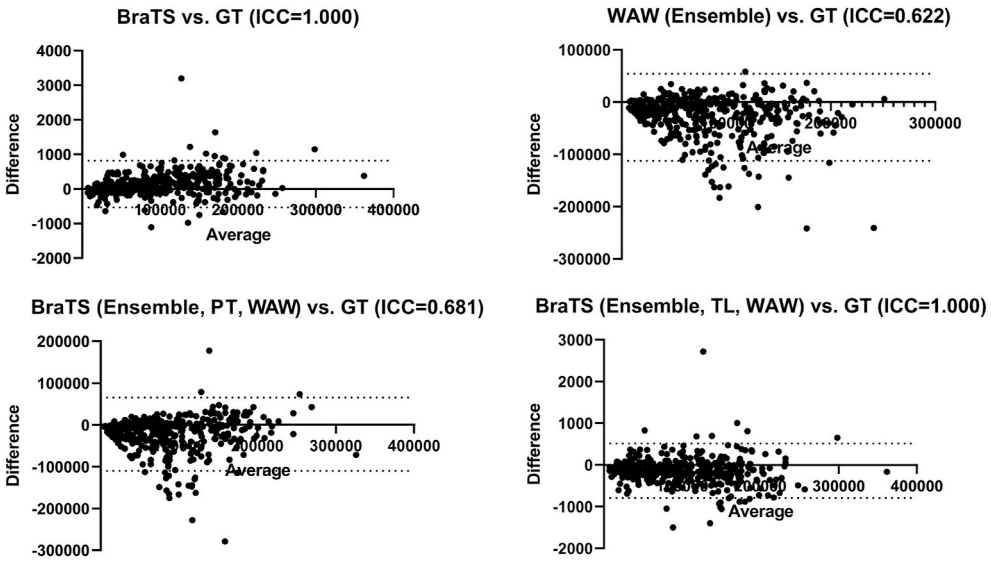

Fig. 12. Bland-Altman plots for the nnU-Net models trained in different settings over all training BraTS patients showing the agreement between the automatically extracted and ground-truth (GT) WT volume, quantified using the Intraclass Correlation Coefficient (ICC).

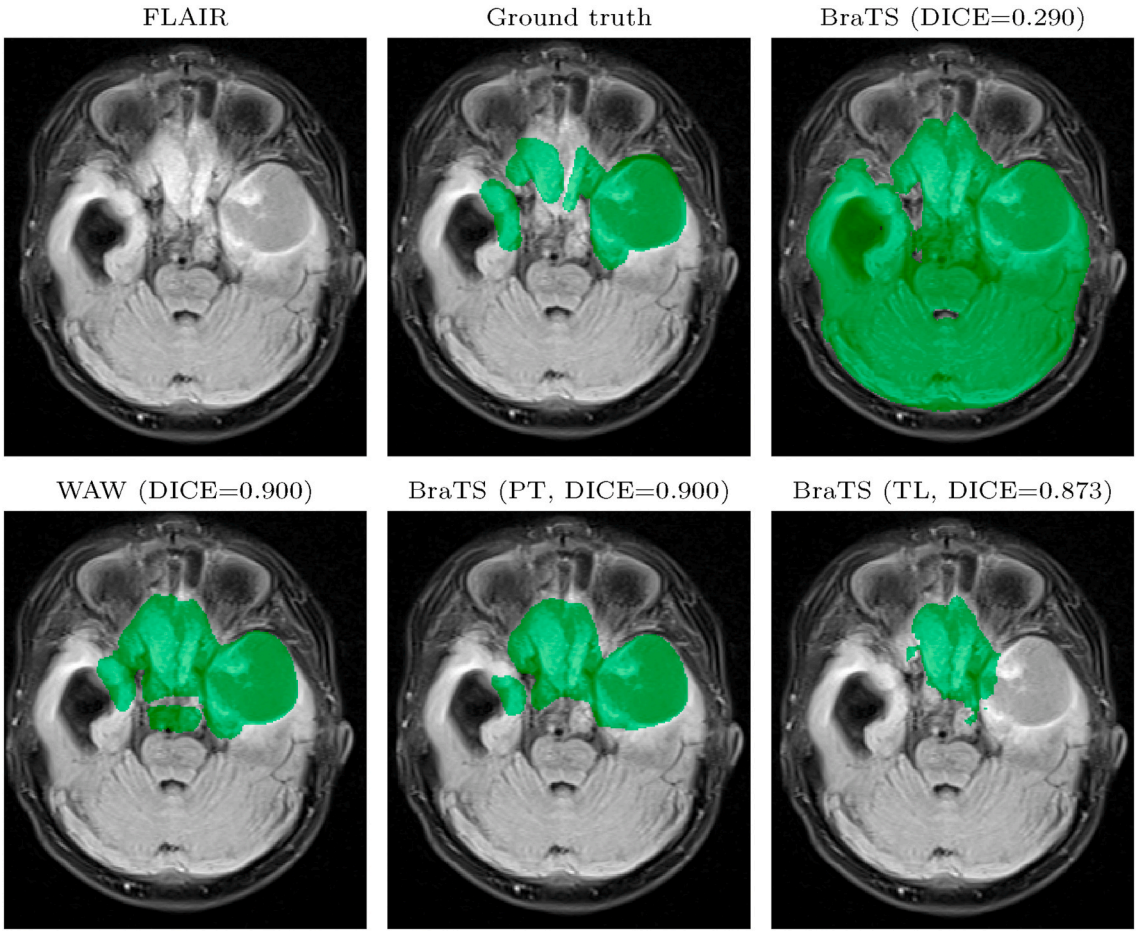

Fig. 13. Segmentation of a test WAW patient (Patient 15). For this MRI, the nnU-Net model trained on BraTS (Tr) delivered the largest DICE score across all WAW patients (i.e., the segmentation quality obtained for all other test WAW patients using this model was worse). Note that other deep networks obtained much better segmentation.

capable of detecting not only OPGs, but also other types of brain lesions. Additionally, training separate nnU-Nets over subsets of available ground-truth MRIs and combining them into ensemble segmentation engines that aggregate base models showed that the worst-performing nnU-Nets are effectively compensated using the outcomes of other deep networks, and significantly improved all quality metrics quantifying the performance of base models. It is, however, interesting to observe one exception in Fig. 11—albeit the distribution of the H95

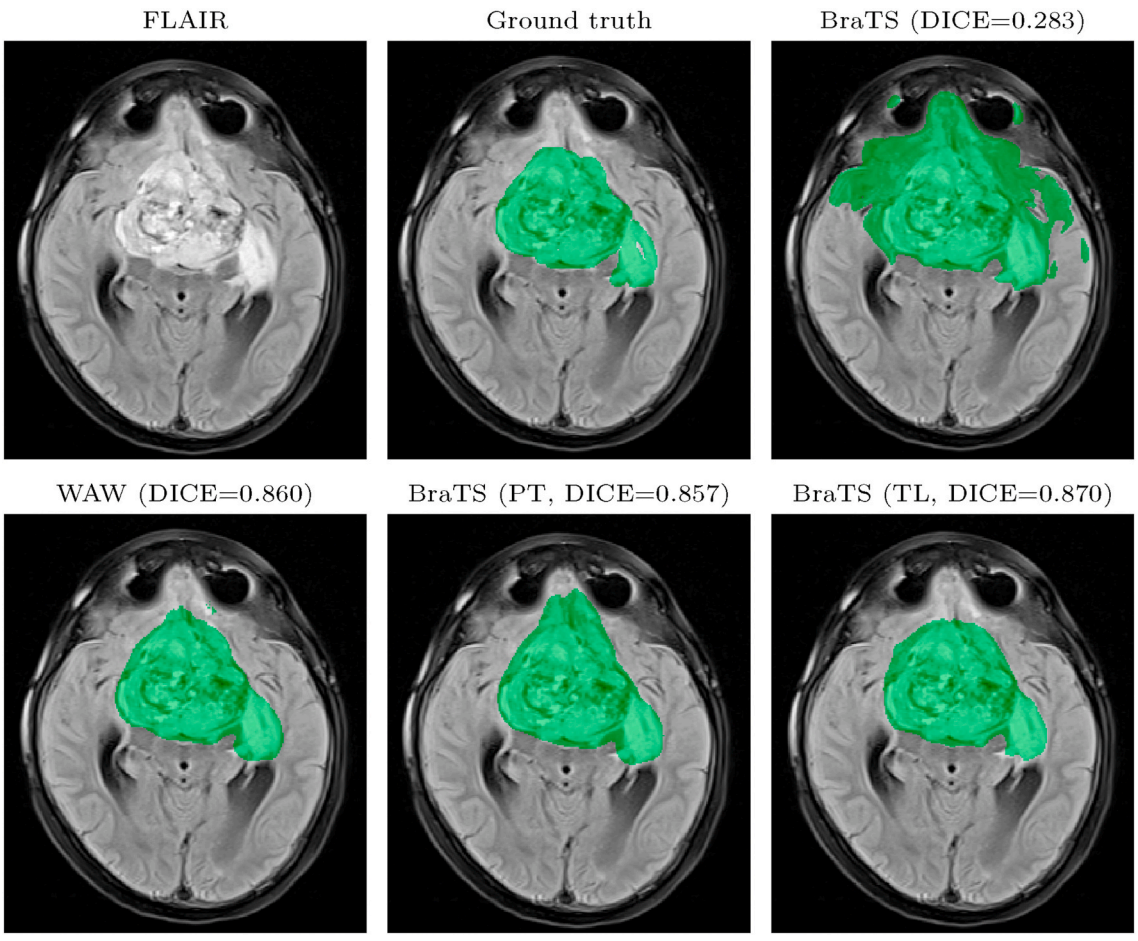

Fig. 14. Segmentation of a test WAW_{Test} patient (Patient 22). For this MRI, the nnU-Net model trained on BraTS (Tr) delivered the largest DICE score across all WAW_{Test} patients (i.e., the segmentation quality obtained for all other test WAW_{Test} patients using this model was worse). Note that other deep networks obtained much better segmentation.

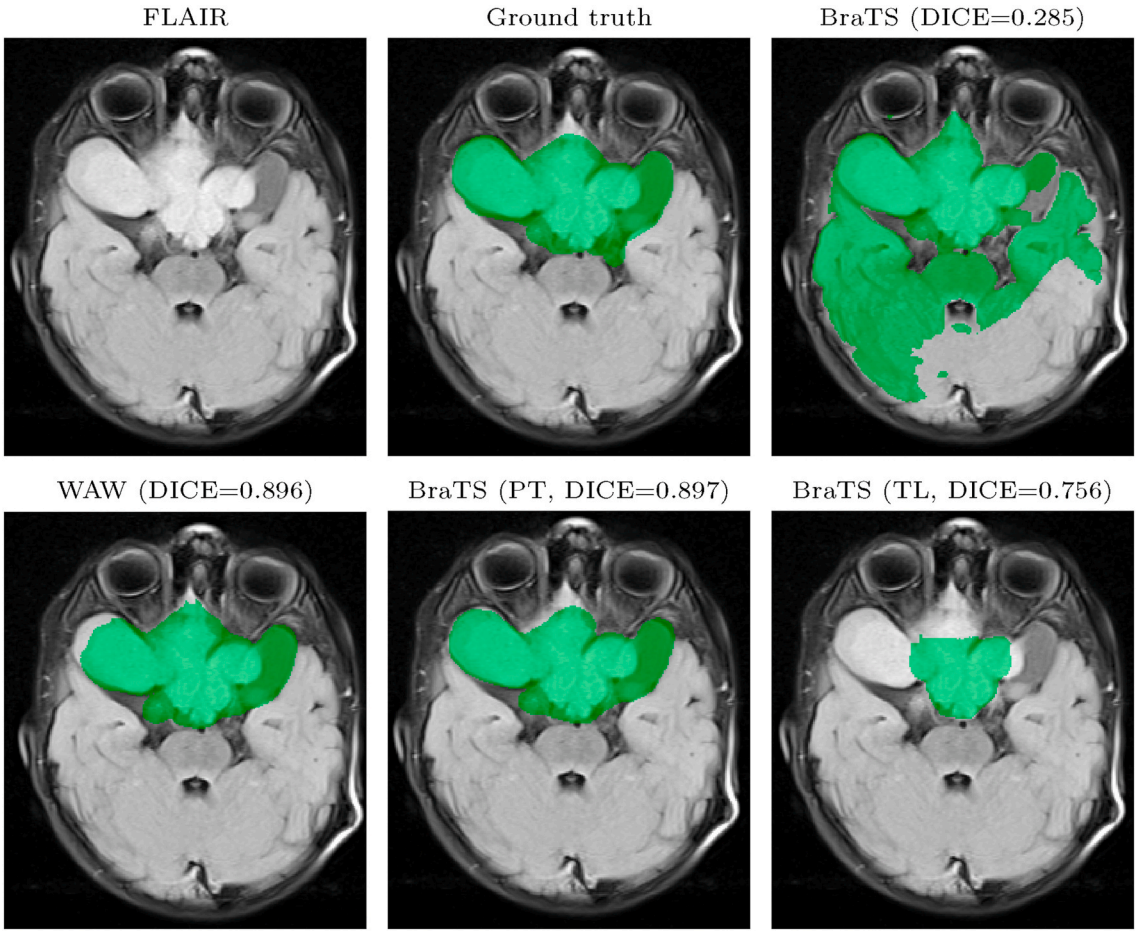

Fig. 15. Segmentation of a test WAW patient (Patient 7). For this MRI, the nnU-Net model trained in the PT strategy delivered the largest DICE score (i.e., the segmentation quality obtained for all other test WAW patients using this model was worse).

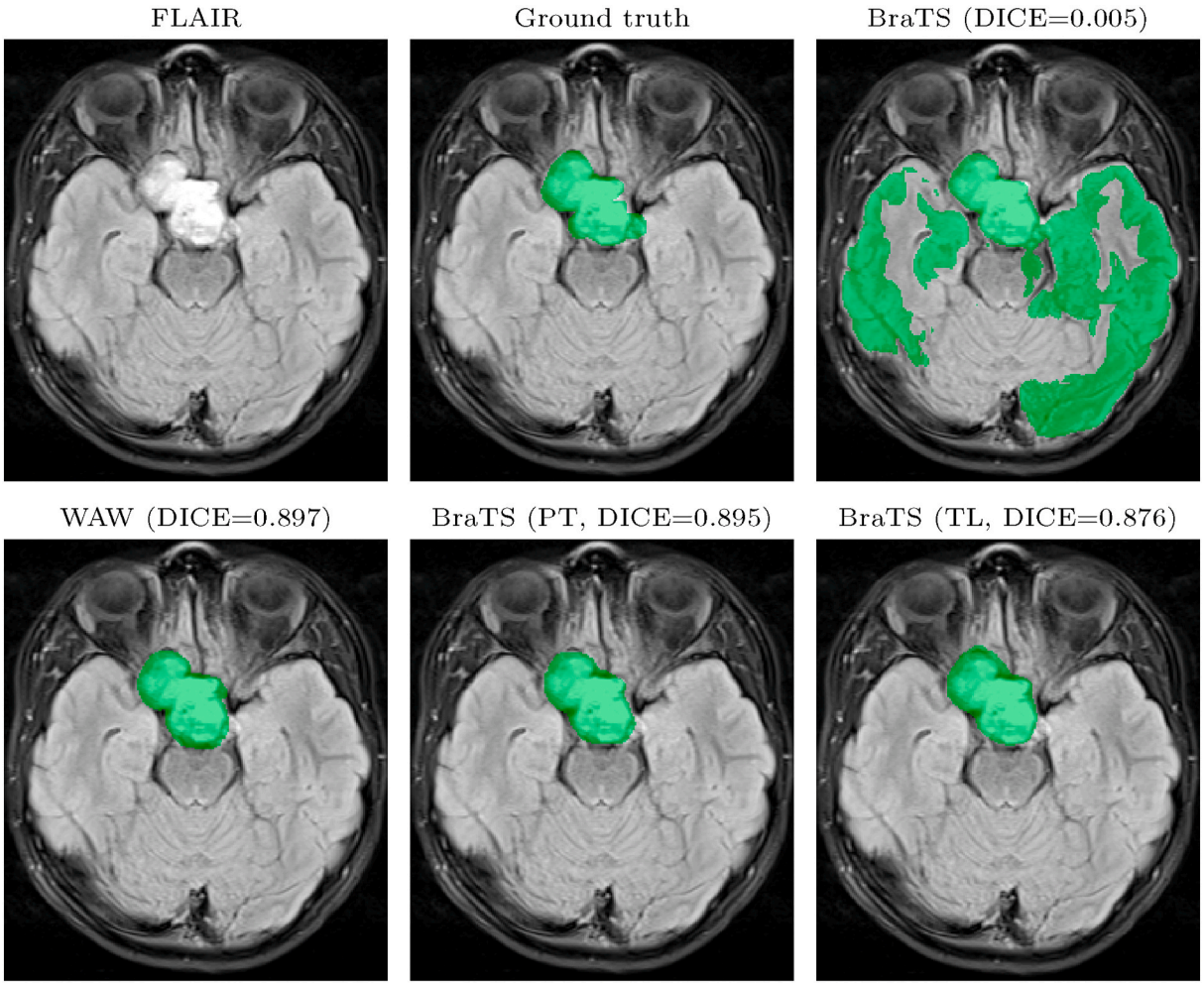

Fig. 16. Segmentation of a test WAW patient (Patient 20). For this MRI, the nnU-Net model trained in the TL strategy delivered the largest DICE score (i.e., the segmentation quality obtained for all other test WAW patients using this model was worse).

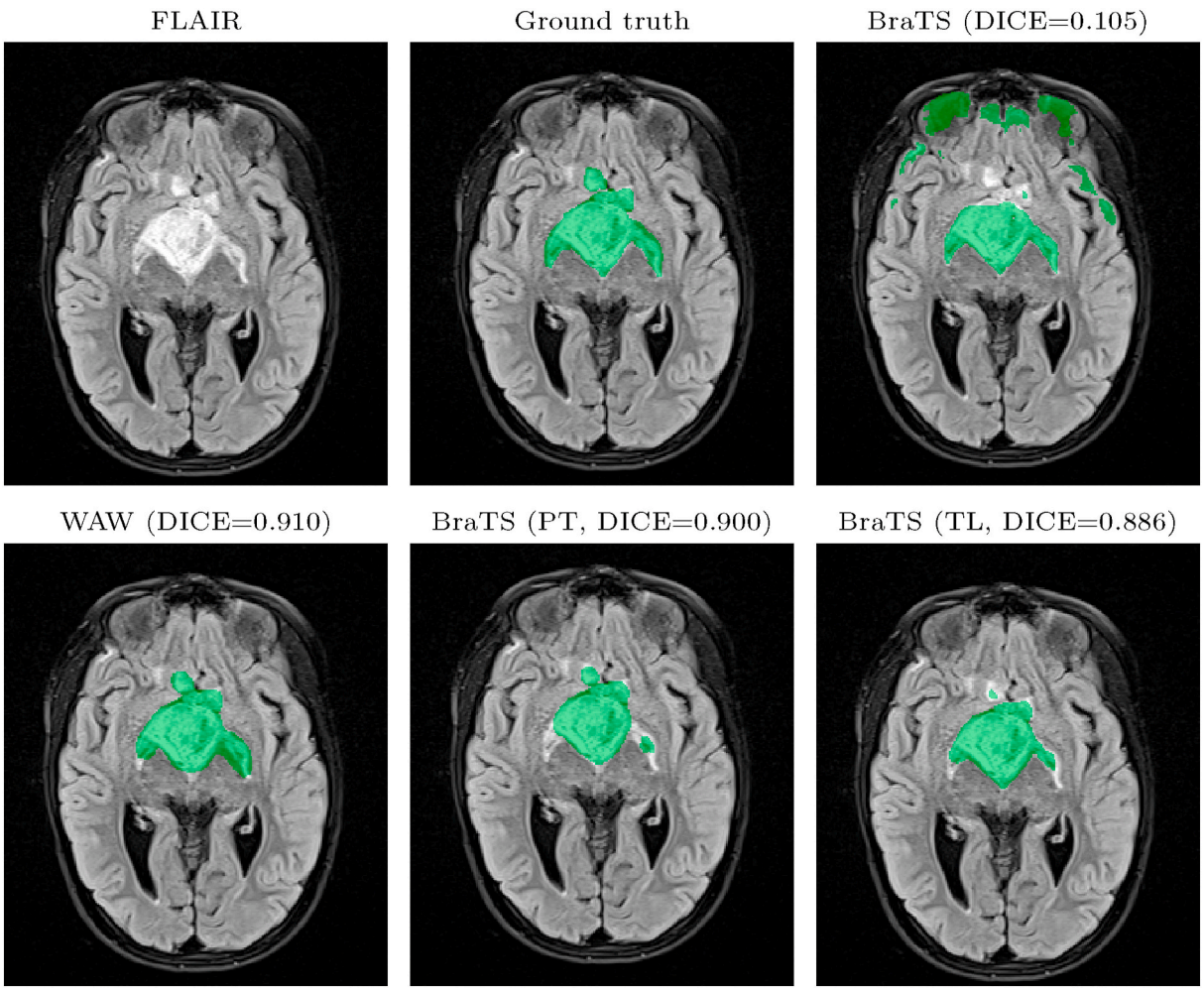

Fig. 17. Segmentation of a test WAW_{Test} patient (Patient 6). For this MRI, all nnU-Net ensembles delivered the largest DICE scores across all WAW_{Test} patients (i.e., the segmentation quality obtained for all other test WAW_{Test} patients using this model was worse). Note that other deep networks obtained much better segmentation.

values was improved for BraTS (Ensemble, TL, WAW) when compared to the base models, we can notice a whiskers indicating an outlying segmentation with a large H95 value. It is attributed to the fact that the ensemble failed to segment a single patient belonging to BraTS (V), and did not annotate any voxels as tumorous, whereas the base models delivered either FP voxels or the DICE value being very close (but not equal to) zero. If the segmentation mask submitted for assessment does not include any tumorous voxels and the corresponding ground-truth manual segmentation is not empty, then the validation server returns the maximal H95 value of approximately 371 mm, being the length of the diagonal of an input MRI ($240 \times 240 \times 155$, with the 1 mm^3 voxel size)—this is the outlying value which can be observed for BraTS (Ensemble, TL, WAW) in this case.

The qualitative examples gathered in Fig. 18 visualize the best segmentations (according to the DICE scores) obtained for the BraTS (V) MRIs using nnU-Net trained over BraTS (Tr), and using all investigated ensembles. It is interesting to note that the results presented for a selected MRI frame of Patient 70 may indicate a high-quality tumor delineation obtained by BraTS (Ensemble, TL, WAW), whereas the corresponding DICE amounted to 0.394 (unsatisfactory segmentation). However, investigating the coronal plane of this MRI in Fig. 19, together with other slices in the image, helps us spot significant discontinuities in the segmented tumor. To provide more robust measurements of the tumor spatial characteristics and tumor burden, volumetric analysis could be performed, as it can deliver more accurate estimation of the tumor's size than the uni- and bidimensional estimations. High agreement between the volumes calculated based on the segmentations obtained using the approaches proposed in this work and the corresponding ground-truth delineations (Fig. 7) suggests that the automated end-to-end analysis pipeline may extract quantitative OPG's volumetric characteristics that correspond to those elaborated by senior radiologists while maintaining full reproducibility of the results, hence leading to more objective analysis of MRI scans.

5. Conclusions

Radiological observation and segmentation of pediatric OPGs is an integral step of efficient management in this type of tumors, as these steps can deliver quantitative measures capturing the current state of the disease. Unfortunately, notable intra- and inter-rater variability in the process of manual delineations of OPGs (and other tumors) from MRI may easily lead to designing sub-optimal treatment pathways. To ensure full reproducibility of the segmentation process, we proposed a deep learning algorithm for the task of OPG detection and segmentation, which is built upon the recent advances in the machine learning field. We suggested the training strategies that can be helpful in elaborating the deep models capable of delineating not only OPGs, but also other brain lesions, such as low- and high-grade gliomas in adults. Additionally, we showed that building segmentation ensembles leads to the combined models that effectively compensate the worst-performing

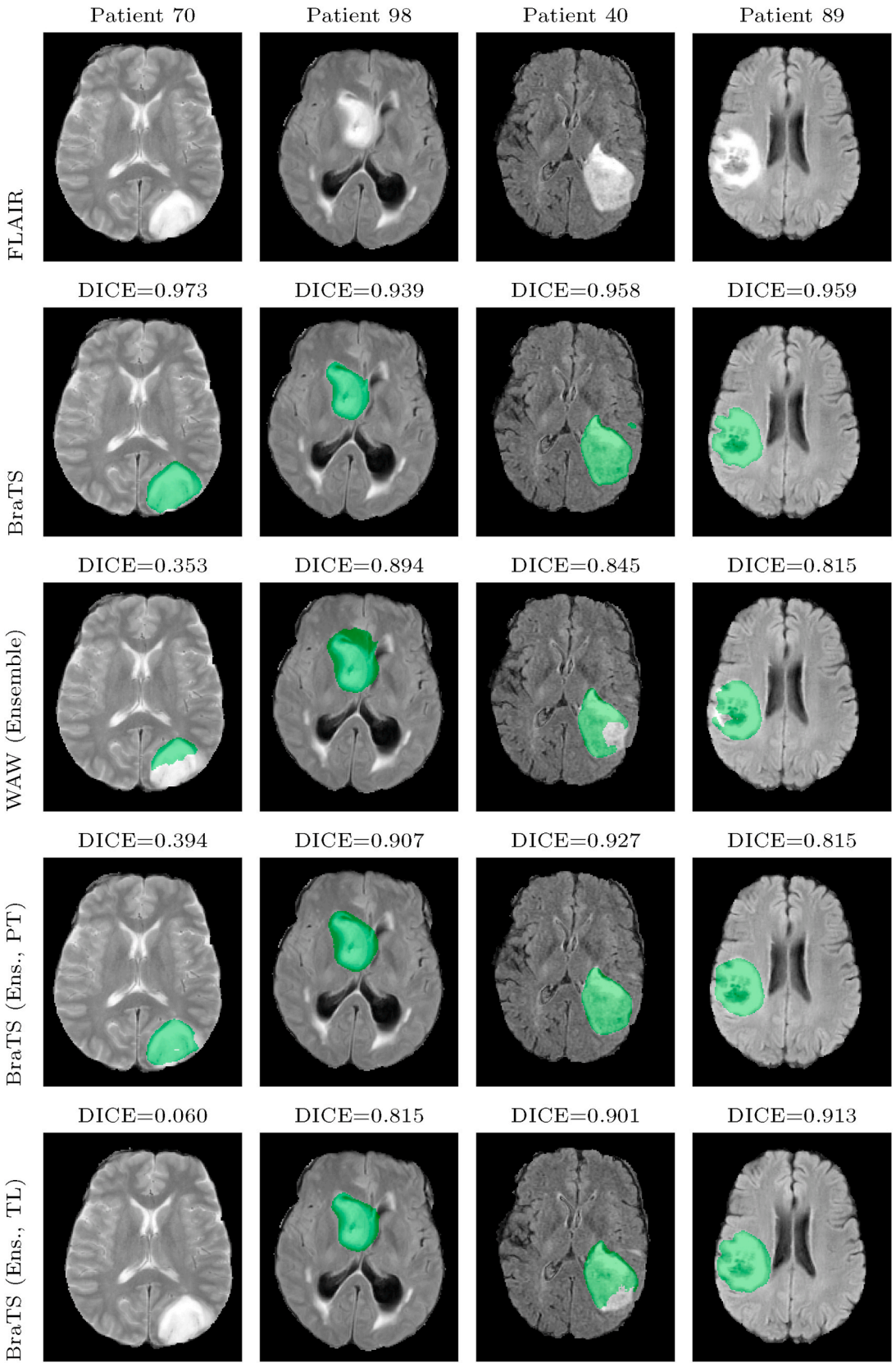

Fig. 18. Segmentations of the validation BraTS patients, for which the nnU-Net model trained solely on the training BraTS dataset (Patient 70), and the nnU-Net ensembles trained over WAW (Patient 98), and in the PT (Patient 40) and TL (Patient 89) strategies, delivered the largest DICE scores. For brevity, we use short names of the ensembles.

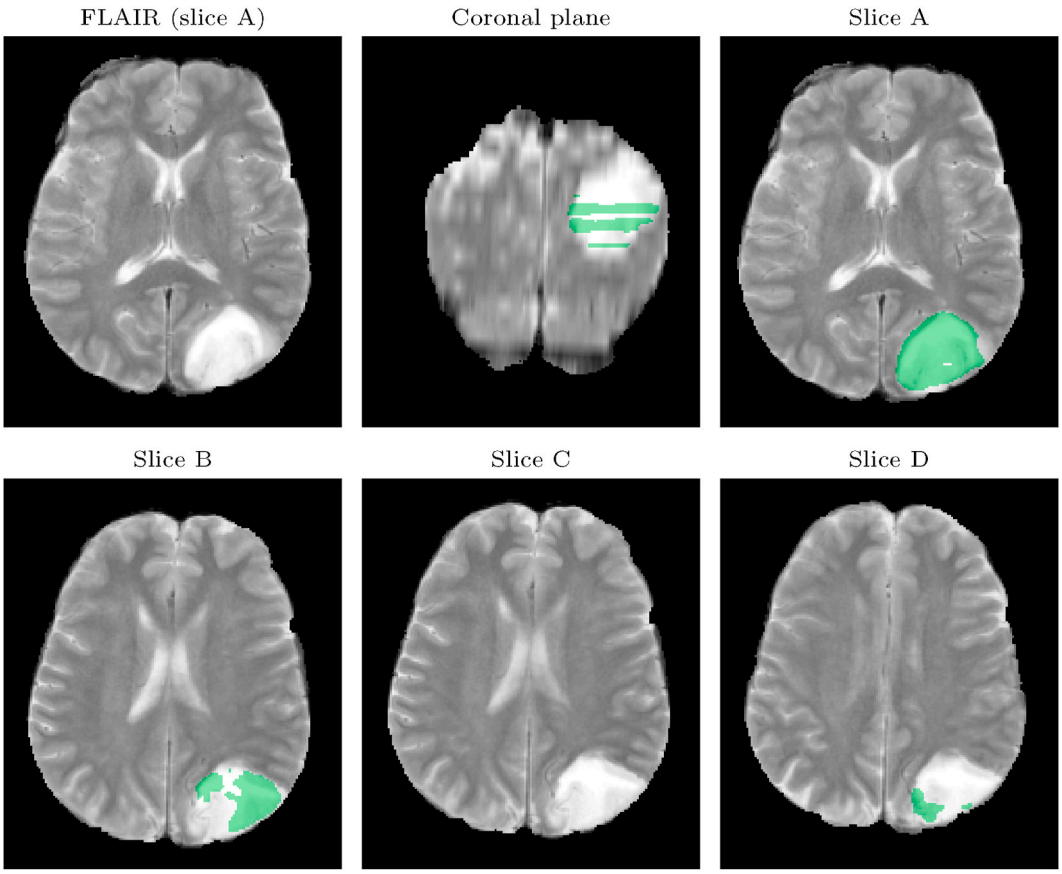

Fig. 19. Qualitative analysis of the segmentation results may easily be misleading if performed over 2D slices. The nnU-Net model trained in the PT strategy delivered a high-quality WT mask in Slice A for this scan (Patient 70), but it failed in other slices of the same scan (B–D). Overall, the DICE score amounted to 0.394 in this case. In the coronal view, we can observe that this model was able to detect a lesion only in selected slices, hence the automatically calculated tumor volume would be much smaller than it should be, and the uni- and bidimensional automatic measurements would fail too.

base classifiers through aggregating the outcomes of all models included in an ensemble. Our rigorous experimental validation, performed over the clinically acquired OPG MRI data alongside the publicly available dataset of LGG/HGG gliomas showed that the proposed models are able to accurately delineate OPGs that manifest various appearance and location characteristics. The qualitative analysis revealed that the OPG volumes extracted using our techniques are in agreement with manual calculations delivered by senior radiologists. Finally, the experiments showed that the proposed techniques offer real-time operation, and the average end-to-end analysis time of a single MRI was less than 10 s, indicating its possibility of incorporating it into clinical workflows without negatively affecting the overall time of radiological procedures. Although the processing chain still requires clinical validation over more heterogeneous MRI data, we believe that introducing an automated OPG delineation procedure into the OPG management and treatment would allow us to design more objective treatment pathways for such patients. Additionally, as the longitudinal evaluation of tumor size is key in diagnostics, making the delineation process fully reproducible and free from human errors through deploying automated OPG segmentation can help physicians quantitatively track the disease and ultimately improve patient's care.

The results reported in this work are an interesting departure point for further research. Although we quantitatively and qualitatively presented the abilities of the proposed nnU-Net-based techniques over the OPG MRI data, it would be interesting to include OPG MRIs acquired using different scanners, and perhaps other acquisition protocols, to fully understand the generalization capabilities of our framework—the limited heterogeneity of the OPG data can be considered as the most important limitation of the work reported here. Such efforts could lead us to releasing a large-scale MRI benchmark dataset[6] specifically targeting pediatric OPGs which might be ultimately used to confront the existent and emerging algorithms for this task. Larger and more representative datasets would allow us to build efficient segmentation ensembles, as they clearly improve the abilities of base models (here, selecting an appropriate content and size of such ensembles, together with elaborating the most promising aggregation techniques should be explored [40]). Finally, we are currently working on the implementation of the proposed algorithms that could be integrated and used in a hands-free manner over a larger set of MRIs in the clinical workflow [13], and on robustifying the algortihms with respect to missing MRI sequences [41].

Declaration of competing interest

Jakub Nalepa: None Declared.
Szymon Adamski: None Declared.
Krzysztof Kotowski: None Declared.
Sylwia Chelstowska: None Declared.
Magdalena Machnikowska-Sokolowska: None Declared.
Oskar Bozek: None Declared.
Agata Wisz: None Declared.
Elzbieta Jurkiewicz: None Declared.

[6] Such OPG set could be made publicly available through various initiatives, such as the Medical Segmentation Decathlon (http://medicaldecathlon.com/; accessed on October 5, 2021), in order to allow other research groups benefit from such image data.

Acknowledgements

JN was partially supported by the Silesian University of Technology grant for maintaining and developing research potential. The authors would like to thank Marek Pitura (Future Processing Healthcare) for his valuable help in managing this study.

This paper is in memory of Dr. Grzegorz Nalepa, an extraordinary scientist and pediatric hematologist/oncologist at Riley Hospital for Children, Indianapolis, USA, who helped countless patients and their families through some of the most challenging moments of their lives.

References

[1] I. Fried, U. Tabori, T. Tihan, A. Reginald, E. Bouffet, Optic pathway gliomas: a review, CNS oncology 2 (2) (2013) 143–159.
[2] N. Rasool, J.G. Odel, M. Kazim, Optic pathway glioma of childhood, Curr. Opin. Ophthalmol. 28 (3) (2017).
[3] R.E. Friedrich, M.A. Nuding, Optic pathway glioma and cerebral focal abnormal signal intensity in patients with neurofibromatosis type 1: characteristics, treatment choices and follow-up in 134 affected individuals and a brief review of the literature, Anticancer Res. 36 (8) (2016) 4095–4121.
[4] V. Robert-Boire, L. Rosca, Y. Samson, L.H. Ospina, S. Perreault, Clinical presentation and outcome of patients with optic pathway glioma, Pediatr. Neurol. 75 (2017) 55–60.
[5] E. Trevisson, M. Cassina, E. Opocher, V. Vicenzi, M. Lucchetta, R. Parrozzani, G. Miglionico, R. Mardari, E. Viscardi, E. Midena, M. Clementi, Natural history of optic pathway gliomas in a cohort of unselected patients affected by neurofibromatosis 1, J. Neuro Oncol. 134 (2) (2017) 279–287.
[6] W.S. Müller-Forell, K. Sartor, Imaging of Orbital and Visual Pathway Pathology, Springer, 2005.
[7] R.K. Imes, W.F. Hoyt, Magnetic resonance imaging signs of optic nerve gliomas in neurofibromatosis 1, Am. J. Ophthalmol. 111 (6) (1991) 729–734.
[8] M.J. Binning, J.K. Liu, J.R.W. Kestle, D.L. Brockmeyer, M.L. Walker, Optic pathway gliomas: a review, Neurosurg. Focus 23 (5) (2007) E2.
[9] R. Listernick, R.E. Ferner, G.T. Liu, D.H. Gutmann, Optic pathway gliomas in neurofibromatosis-1: controversies and recommendations, Ann. Neurol. 61 (3) (2007) 189–198.
[10] P.Y. Wen, D.R. Macdonald, D.A. Reardon, T.F. Cloughesy, A.G. Sorensen, E. Galanis, J. Degroot, W. Wick, M.R. Gilbert, A.B. Lassman, C. Tsien, T. Mikkelsen, E.T. Wong, M.C. Chamberlain, R. Stupp, K.R. Lamborn, M.A. Vogelbaum, M.J. van den Bent, S.M. Chang, Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group, J. Clin. Oncol. 28 (11) (2010) 1963–1972.
[11] K. Chang, A.L. Beers, H.X. Bai, J.M. Brown, K.I. Ly, X. Li, J.T. Senders, V. K. Kavouridis, A. Boaro, C. Su, W.L. Bi, O. Rapalino, W. Liao, Q. Shen, H. Zhou, B. Xiao, Y. Wang, P.J. Zhang, M.C. Pinho, P.Y. Wen, T.T. Batchelor, J.L. Boxerman, O. Arnaout, B.R. Rosen, E.R. Gerstner, L. Yang, R.Y. Huang, J. Kalpathy-Cramer, Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement, Neuro Oncol. 21 (11) (2019) 1412–1422.
[12] URL S. Bakas, M. Reyes, et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, Apr. 2019. arXiv:1811.02629 [cs, stat]ArXiv: 1811.02629, http://arxiv.org/abs/1811.02629.
[13] J. Nalepa, P. Ribalta Lorenzo, M. Marcinkiewicz, B. Bobek-Billewicz, P. Wawrzyniak, M. Walczak, M. Kawulok, W. Dudzik, K. Kotowski, I. Burda, B. Machura, G. Mrukwa, P. Ulrych, M.P. Hayball, Fully-automated deep learning-powered system for DCE-MRI analysis of brain tumors, Artif. Intell. Med. 102 (2020) 101769.
[14] K. Kotowski, S. Adamski, W. Malara, B. Machura, L. Zarudzki, J. Nalepa, Segmenting brain tumors from MRI using cascaded 3D U-nets, in: A. Crimi, S. Bakas (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 265–277.
[15] B.H. Menze, A. Jakab, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imag. 34 (10) (2015) 1993–2024.
[16] URL U. Baid, S. Ghodasara, et al., The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, Jul. 2021. arXiv: 2107.02314 [cs]ArXiv: 2107.02314, http://arxiv.org/abs/2107.02314.
[17] J. Liu, M. Li, J. Wang, F. Wu, T. Liu, Y. Pan, A survey of MRI-based brain tumor segmentation methods, Tsinghua Sci. Technol. 19 (6) (2014) 578–595.
[18] A. Wadhwa, A. Bhardwaj, V. Singh Verma, A review on brain tumor segmentation of MRI images, Magn. Reson. Imag. 61 (2019) 247–259.
[19] M.T.M. Park, J. Pipitone, L.H. Baer, J.L. Winterburn, Y. Shah, S. Chavez, M. M. Schira, N.J. Lobaugh, J.P. Lerch, A.N. Voineskos, M.M. Chakravarty, Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates, Neuroimage 95 (2014) 217–231.
[20] A.I. Poernama, I. Soesanti, O. Wahyunggoro, Feature extraction and feature selection methods in classification of brain MRI images: a review, in: Proc. IEEE IBITeC, vol. 1, 2019, pp. 58–63.
[21] H.K. Abbas, N.A. Fatah, H.J. Mohamad, A.A. Alzuky, Brain tumor classification using texture feature extraction, J. Phys. Conf. 1892 (1) (2021), 012012.
[22] M.A. Naser, M.J. Deen, Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images, Comput. Biol. Med. 121 (2020) 103758.
[23] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[24] H. Saleem, A.R. Shahid, B. Raza, Visual interpretability in 3D brain tumor segmentation network, Comput. Biol. Med. 133 (2021) 104410.
[25] M. Artzi, S. Gershov, L. Ben-Sira, J. Roth, D. Kozyrev, B. Shofty, T. Gazit, T. Halag-Milo, S. Constantini, D.B. Bashat, Automatic segmentation, classification, and follow-up of optic pathway gliomas using deep learning and fuzzy c-means clustering based on MRI, Med. Phys. 47 (11) (2020) 5693–5701.
[26] L. Weizman, L. Joskowicz, L. Ben-Sira, R. Precel, D. Ben-Bashat, Automatic segmentation of optic pathway gliomas in MRI, in: Proc, IEEE ISBI, 2010, pp. 920–923.
[27] B. Shofty, L. Weizman, L. Joskowicz, S. Constantini, A. Kesler, D. Ben-Bashat, M. Yalon, R. Dvir, S. Freedman, J. Roth, L. Ben-Sira, MRI internal segmentation of optic pathway gliomas: clinical implementation of a novel algorithm, Child's Nerv. Syst. 27 (8) (2011) 1265–1272.
[28] L. Weizman, L. Ben Sira, L. Joskowicz, S. Constantini, R. Precel, B. Shofty, D. Ben Bashat, Automatic segmentation, internal classification, and follow-up of optic pathway gliomas in MRI, Med. Image Anal. 16 (1) (2012) 177–188.
[29] B. Park, W.R. Windham, K.C. Lawrence, D.P. Smith, Classification of hyperspectral imagery for identifying fecal and ingesta contaminants, in: Monitoring Food Safety, Agriculture, and Plant Health, vol. 5271, International Society for Optics and Photonics, 2004, pp. 118–127.
[30] A. Mansoor, I. Li, R.J. Packer, R.A. Avery, M.G. Linguraru, Joint deep shape and appearance learning: application to optic pathway glioma segmentation, in: Medical Imaging 2017: Computer-Aided Diagnosis, vol. 10134, International Society for Optics and Photonics, 2017, p. 101341O.
[31] O. Ronneberger, P. Fischer, T. Brox, U-Net, Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention, Springer International Publishing, Cham, 2015, pp. 234–241.
[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[33] F. Isensee, P.F. Jäger, P.M. Full, P. Vollmuth, K.H. Maier-Hein, nnU-Net for brain tumor segmentation, in: A. Crimi, S. Bakas (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2021, pp. 118–132.
[34] T. Rohlfing, N.M. Zahr, E.V. Sullivan, A. Pfefferbaum, The SRI24 multichannel atlas of normal adult human brain structure, Hum. Brain Mapp. 31 (5) (2010) 798–819.
[35] S. Haller, E. Kövari, F.R. Herrmann, V. Cuvinciuc, A.-M. Tomm, G.B. Zulian, K.-O. Lovblad, P. Giannakopoulos, C. Bouras, Do brain T2/FLAIR white matter hyperintensities correspond to myelin loss in normal aging? a radiologic-neuropathologic correlation study, Acta Neuropathologica Communications 1 (1) (2013) 14.
[36] Y. Wen, L. Chen, Y. Deng, C. Zhou, Rethinking pre-training on medical imaging, J. Vis. Commun. Image Represent. 78 (2021) 103145.
[37] J. Nalepa, M. Myller, M. Kawulok, Transfer learning for segmenting dimensionally reduced hyperspectral images, Geosci. Rem. Sens. Lett. IEEE 17 (7) (2020) 1228–1232.
[38] J. Nalepa, M. Marcinkiewicz, M. Kawulok, Data augmentation for brain-tumor segmentation: a review, Front. Comput. Neurosci. 13 (2019) 83.
[39] S. Jadon, A survey of loss functions for semantic segmentation, Proc. IEEE CIBCB (2020) 1–5.
[40] P. Bosowski, J. Bosowska, J. Nalepa, Evolving deep ensembles for detecting Covid-19 in chest X-Rays, in: Proc, IEEE ICIP, 2021, pp. 3772–3776.
[41] Y. Shen, M. Gao, Brain tumor segmentation on MRI with missing modalities, in: A. C.S. Chung, J.C. Gee, P.A. Yushkevich, S. Bao (Eds.), Proc. ICMI, Springer International Publishing, Cham, 2019, pp. 417–428.