Brain Tumor Segmentation from Multiparametric MRI Using a Multi-encoder U-Net Architecture

Saruar Alam[1,2(✉)], Bharath Halandur[2,3], P. G. L. Porta Mana[2,4], Dorota Goplen[5], Arvid Lundervold[1,2], and Alexander Selvikvåg Lundervold[2,4]

[1] Department of Biomedicine, University of Bergen, Bergen, Norway
saruar.alam@uib.no
[2] Mohn Medical Imaging and Visualization Centre (MMIV), Department of Radiology, Haukeland University Hospital, Bergen, Norway
[3] Department of Clinical Science, University of Bergen, Bergen, Norway
[4] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway
[5] Department of Oncology, Haukeland University Hospital, Bergen, Norway

Abstract. This paper describes our submission to Task 1 of the RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2021, where the goal is to segment brain glioblastoma sub-regions in multi-parametric MRI scans. Glioblastoma patients have a very high mortality rate; robust and precise segmentation of the whole tumor, tumor core, and enhancing tumor subregions plays a vital role in patient management. We design a novel multi-encoder, shared decoder U-Net architecture aimed at reducing the effect of signal artefacts that can appear in single channels of the MRI recordings. We train multiple such models on the training images made available from the challenge organizers, collected from 1251 subjects. The ensemble-model achieves Dice Scores of 0.9274 ± 0.0930, 0.8717 ± 0.2456, and 0.8750 ± 0.1798; and Hausdorff distances of 4.77 ± 17.05, 17.97 ± 71.54, and 10.66 ± 55.52; for whole tumor, tumor core, and enhancing tumor, respectively; on the 570 test subjects assessed by the organizer. We investigate the robustness of our automated segmentation system and discuss its possible relevance to existing and future clinical workflows for tumor evaluation and radiation therapy planning.

Keywords: Brain tumor segmentation · CNNs · BraTS 2021

1 Introduction

A primary brain tumor typically originates from glial cells, while a secondary (metastatic) brain tumor has its origin in another organ. Gliomas are the most

This work was supported by the Trond Mohn Research Foundation [grant number BFS2018TMT07]. Data used in this publication were obtained as part of the RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge project through Synapse ID (syn25829067).

A. Crimi and S. Bakas (Eds.): BrainLes 2021, LNCS 12963, pp. 289–301, 2022.
https://doi.org/10.1007/978-3-031-09002-8_26

common type of primary brain tumor. Based on their lethality they are classified into low-grade (LGG) and high-grade (HGG). LGGs are less malignant and grow slowly, whereas the more lethal HGGs are highly malignant and grow rapidly. Accurate characterization and localization of tumor-tissue types plays a key role in brain-tumor diagnosis and patient management. Neuroimaging methods, in particular magnetic resonance imaging (MRI), provide anatomical and pathophysiological information about brain tumors and can aid in diagnosis, treatment planning and follow-up of patients. Manual segmentation of tumor tissue is a tedious and time and resource consuming task, prone to inter and intra-rater variability.

The objective of the present work is to develop a model that can aid the clinicians in distinguishing between tumor and normal brain tissue. A reliable tumor segmentation system may have enormous impact both on preoperative tumor staging and planning of the resection extent and possibly increase/decrease the rate of gross tumor resection, the postoperative assessment of residual tumor, and on radiotherapy planning. Automatic segmentation may be applied for delineating the critical structures as well as residual tumor and risk zones. Such a system may contribute to more standardized treatment and limit the radiotherapy dose to critical structures.

Deep learning models most frequently used for image segmentation have encoder-decoder architectures based on convolutional neural networks [13]. A particular example is U-Net [20]. Its basic architecture has skip connections, but has recently been improved with several additional features such as residual blocks [7], attention mechanisms [17], and squeeze-and-excitation blocks [8]. Making such models sufficiently robust to signal artifacts present in MRI channels is an important and challenging goal. We propose a novel U-Net architecture with five encoders and a shared decoder with attention blocks. Our architectural modifications to the standard U-Net [20] are intended to reduce the impact of signal artifacts when segmenting brain tumor subregions.

2 Dataset

We use preprocessed multiparametric MRI recordings from the data collection provided by the Brain Tumor Segmentation challenge 2021 (BraTS 2021) [1–5]. This dataset contains 2000 multiparametric examinations, split into three sets: 1251 for training, 219 for validation, 530 for testing. Each subject's examination consists of four MRI sequences: T1w, T1cw, T2w, and FLAIR. All MRI recordings were preprocessed by the organizer through (i) DICOM to NIFTI file format conversion; (ii) re-orientation and co-registering to a fixed anatomical template (SRI24) [19] to obtain an isotropic spatial resolution of $1 \times 1 \times 1$ mm^3 and matrix size of $240 \times 240 \times 155$; (iii) skull-stripping.

Each subject has three mutually exclusive delineated regions: necrotic and non-enhancing tumor core (NCR/NET; intensity label 1), peritumoral edema (ED; intensity label 2), and enhancing tumor (ET; intensity label 4) stored in a 3D segmentation mask image (ground truth) with the same spatial resolution

and matrix size as the corresponding channel images. The organizer consider three non-mutually-exclusive tumoral regions to standardize overall assessment and evaluation in the competition: (i) enhancing tumor (ET); (ii) tumor core (TC = ET ∪ NCR), (iii) whole tumor (WT = ET ∪ NCR ∪ ED). The three labeled regions ET, TC, WT are clinically relevant to patient management (e.g. treatment, prognosis). The organizer's annotation procedure followed two steps: (i) fusion of tumor regions of top-performing segmentation models of the previous editions of the challenge (DeepMedic [12], DeepScan [15], and nnUNet [9]), (ii) manual refinement of the fused tumor regions by a team of neuroradiologists.

The ground truth segmentations of the 1251 patient examinations used for training are available to the BraTS 2021 participants. The ground truths for the validation and testing subjects are withheld from the participants. For the validation data, the participants get access to the images without masks and are tasked with producing masks that can be uploaded to the competition's evaluation platform. For the test phase the participants do not get access to the test data, but must instead upload their algorithms to the evaluation platform. This prevents overfitting of the models to the test data that could be caused by multiple iterations of submission, evaluation and model development.

3 Methods

3.1 Data Splitting Approach

Gliomas are heterogeneous, varying in size, shape, volume, and location in the brain. Splitting the data at random into training and validation sets might lead to under- or over-representation of some volume subranges in either set, with a consequent bias in the model's learning and our estimate of its generalization performance. To avoid such a bias the data are first binned into equal ET-volume subranges, and a 75:25 random split is then performed bin-wise to construct a training set and a training-validation set of 923 and 328 subjects, respectively. We choose to use the ET volume for binning because this region is more difficult to detect than the WT and TC regions.

3.2 Normalization and Data Augmentation

The BraTS challenge's dataset is acquired from several institutions worldwide having different scanners and acquisition protocols. This heterogeneity may lead to a spurious non-uniform intensity distribution in the dataset (domain shift), to which a deep learning model may not be robust. To combat this we employ histogram standardization [16], i.e. the centering and standardization of the voxel intensity distribution, on each MRI scan of the training, training-validation, and validation sets.

Data augmentation, consisting in feeding the model with additional transformed views of the same image, is a crucial training step. It provides the model

with some degree of invariance under particular natural transformations, improving its generalization ability. The following nine augmentation transforms are applied on the fly during training:

1. One of the following transforms two at each iteration are selected with a probability of 0.4:
   i) Affine transformation (scale range: 0.9 to 1.1; rotation range: −15 to +15) with a probability of 0.6
   ii) Elastic deformation (number of control points: 7; displacement: 0 to 7, in a coarse grid) with a probability of 0.4
2. Crop of the foreground area and then extraction of a $128 \times 128 \times 128$ patch by considering a ratio of non-tumor and tumor region
3. Zoom with minimum factor 0.9 to maximum 1.2 with a probability of 0.15
4. Additive i.i.d. Gaussian $N(0, 0.01)$ noise with a probability of 0.15
5. Gaussian smoothing with $\sigma \in [0.5, 1.15]$ with a probability of 0.15
6. Shift of intensity with a factor of 0.3 with a probability of 0.15
7. Rescaling of intensity with a factor of 0.3 with a probability of 0.15
8. Addition of Gibbs noise with $\alpha \in [0.1, 0.5]$ with a probability of 0.15
9. Individual flip of each image axis with a probability of 0.5

For histogram standardization, affine transformation, and elastic deformation we use the implementations in the TorchIO library [18], while the remaining transforms are implemented using the MONAI library [14].

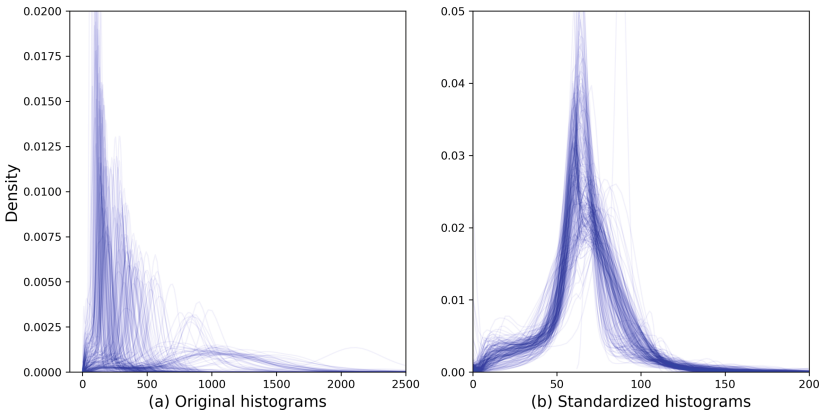

Fig. 1. (a) Original and (b) Standardized histograms of FLAIR images of the organizer's validation set (219 subjects). Note that the signal intensity distributions become more homogeneous across the standardized intensity range.

3.3 Multi-encoder U-Net Architecture

Our proposed architecture is based on an encoder-decoder U-Net, as illustrated in Fig. 2. The encoders reduce the spatial resolution of the feature map while increasing the number of channels at every step. The decoders increase the spatial resolution of the feature map while decreasing the number of channels.

Every encoder branch has five downsampling steps with filter sizes 16, 32, 64, 128, 160, 160. Each step has two consecutive $3 \times 3 \times 3$ convolution layers. Each convolution layer employs an Instance Normalization and a leaky ReLU activation (Conv3d-InstanceNorm3d-LReLU). At the first step of every encoder branch, we use stride size 1 for both convolutional layers. The resolution of the feature map therefore remains the same as the resolution of the input patch at the first step of the encoder. From the second step of the encoder branch to the bottleneck step, the first convolutional layer uses strided convolutions.

Every decoder branch has five upsampling steps, each with two consecutive $3 \times 3 \times 3$ convolution layers preceded by a $2 \times 2 \times 2$ transpose convolution, increasing feature map resolution and reducing channels. The feature map from the corresponding step of the encoder is concatenated with the feature map from the transpose convolution. The concatenated feature map is fed to the two convolutional layers. In the last n layers, a $1 \times 1 \times 1$ convolution finally maps the channels to three tumor regions for the decoder that has n deep-supervision heads (DS heads). The preceding *n-1* DS heads are upsampled with nearest-neighbour interpolation to maintain the same resolution as n-th head. In other words, a model with n heads provides n feature maps having the same resolution as the input patch in training mode. While in inference mode, the model only considers the last layer (or head).

We use five encoders, each training a different MRI-image subset: {T1w, T1cw, T2w, FLAIR}, {T1w, T1cw, FLAIR}, {T1w, T1cw, T2w}, {T1w, T2w, FLAIR}, and {T1cw, T2w, FLAIR}. All encoders share a common decoder. In decoder steps, the feature maps from the corresponding steps of the encoders are concatenated to feed into channel attention blocks. The feature maps from channel attention blocks are concatenated with the feature map from the transpose convolution.

3.4 Training Variants of Multi-encoder U-Net

We employ two different pipelines to train our multiencoder U-Net (MEU-Net) model.

***Pipeline A*:** The models are trained based on three consecutively inclusive BraTS-specific regions: ET $\subseteq$ TC $\subseteq$ WT. Earlier studies [9,10] found that such training increases the model's ability to segment the regions used to rank participants. This pipeline contains two model subtypes: (i) a MEU-Net with four DS heads; (ii) a MEU-Net with three DS heads. In *sub-type (i)*, all DS heads are jointly optimized with a weighted loss. The higher the layer is in the decoder, the larger the weights we assign to the head. We assign larger weights in the upper decoder layer to penalize the higher resolution feature map for achieving better

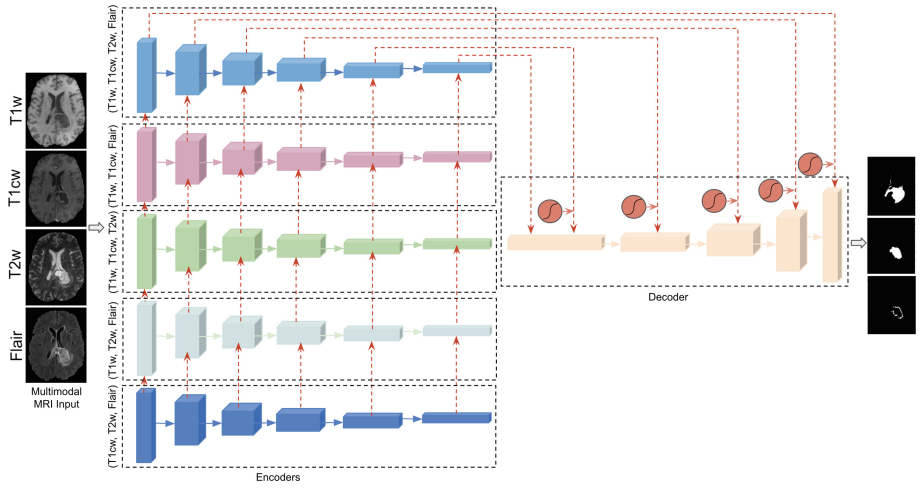

Fig. 2. Adopted multi-encoder-based U-Net architecture

spatial localization. In *sub-type (ii)*, the DS heads are optimized together with a weighted loss. The lower the layer is in the decoder, the larger the weight we provide to the head. Larger weights are provided in lower layers in the decoder to learn from the lower resolution feature map closer to the bottleneck layer, which contains compressed feature representations.

***Pipeline B*:** A MEU-Net is trained on the mutually exclusive BraTS ground truth-specific regions: NCR/NET, ED, ET. The model has four DS heads, optimized with a weighted loss, with larger weights for the upsampling layer of the decoder. This pipeline also contains two model sub-types: (i) a MEU-Net containing encoders similar to the encoders of the models in pipeline A, (ii) a MEU-Net containing encoders with fewer filter set than that of *sub-type (i)*. The *sub-type (ii)* uses fewer filters than *sub-type (i)* to ensure that the model remains within the memory constraints when ensembling.

Pipelines A and B are trained with combined binary cross-entropy and dice loss to optimize each tumor region independently. We use the Ranger21 optimizer with an initial learning rate of $1e-03$ in both pipelines [21]. The Ranger21 combines the AdamW optimizer with various components such as adaptive gradient clipping, gradient centralization, stable weight decay, linear learning rate warm-up, a learning rate scheduler, and lookahead [21]. It has been shown to outperform the Adam optimizer in several benchmark datasets [21]. We observe that the Ranger21 achieves faster convergence and a smooth loss surface during the training phase. For the pipeline A and its subtypes, the MEU-Net is first trained for 120 epochs, using a single split (our training /training-validation set as described in Sect. 3.1). The models are thereafter fine-tuned on the entire training set. For the pipeline B and its subtypes, the MEU-Net is just fine-

tuned on the entire training set using the pre-trained weights from the models of pipeline A.

We employ ensembling to increase the performance of the produced segmentation masks. For the inference on the BratTS validation set, three ensemble sets are employed: (i) *ensemble-1*: the tumor regions WT, TC, ET are obtained by averaging sigmoid outputs of two models available in pipeline A; (ii) *ensemble-2*: the tumor sub-regions NCR/NET, ED, ET are obtained by averaging sigmoid outputs of two models available in pipeline B; (iii) *ensemble-3*: ET is obtained by averaging sigmoid outputs of ET from four different models of pipelines A and B. NCR/NET is obtained by averaging the sigmoid outputs of the complement of the intersection of TC and ET ($(\mathrm{TC} \cap \mathrm{ET})^{\complement}$) from pipeline A, and NCR/NET from pipeline B. Similarly, ED is obtained by averaging the sigmoids outputs of $(\mathrm{WT} \cap (\mathrm{ET} \cup \mathrm{NCR/NET}))^{\complement}$ from pipeline A, and ED from pipeline B.

During inference we perform post-processing to disjoin the overlapping regions and convert them to unique NCR/NET, ED, ET regions in a single multi-class image array, as the latter set is the one required by the competition's evaluation platform. Small tumors are comparatively difficult to segment, and a slight location variance in predicted tumor volume significantly reduces the performance metric. ET has in general a smaller volume than the other regions, and it is particularly challenging to obtain precise segmentations. This also applies to TC in cases where the TC volume is small. As part of our post-processing, we therefore relabeled ET voxels in cases where they number <250 as necrosis (to remain in the TC region).

4 Results

Table 1 shows the result (means, standard deviations, and quartiles of Dice Scores and hausdorff95 distances) of models from pipeline A (PA), Dynamic UNet (DynUNet) [9] and UNEt TRansformer (UNETR) [6]; validated on our internal validation data set. Figure 3 compares the result of pipeline A with that of DynUNet and UNETR (employed from MONAI library [14]). When calculating performance metrics, we provide a full reward (Dice Score = 1 & hausdorff95 distance = 0) to the model for each subject where the model does not predict any tumor and the ground truth also doesn't contain any tumor. The *sub-type (i)* of the pipeline A (MEU-Net(PA,4 Heads)) achieves average Dice Scores of 0.9325, 0.9082, and 0.8863; hausdorff95 distances (HD95) of 6.72, 4.40, and 3.28; for whole tumor, tumor core, and enhancing tumor, respectively. The *sub-type (ii)* of the same pipeline (MEU-Net(PA,3 Heads)) provides similar Dice Scores across the tumor sub-regions. Both MEU-Net(PA,3 Heads) and MEU-Net(PA,4 Heads) perform better than that of DynUNet and UNETR.

Table 2 shows the results for three ensemble models validated on the organizer's validation set. For the WT, TC, ET regions: *ensemble-1* respectively achieves Dice Scores 0.9249, 0.8607, 0.8419, and hausdorff95 distances (HD95) 4.042, 11.40, 17.71; *ensemble-2* achieves Dice Scores 0.9210, 0.8588, 0.8364, and HD95 4.65, 9.96, 17.79; *ensemble-3* achieves Dice Scores 0.9244, 0.8600,

Table 1. Training-validation set results (328 subjects)

Models	Tumor regions	Dice Score					HD95				
		Mean	SD	Q1	Q2	Q3	Mean	SD	Q1	Q2	Q3
DynUNet (3 Heads) [9]	WT	0.9266	0.0774	0.9221	0.9549	0.9702	6.1161	10.2659	1.4142	2.4495	6.0411
	TC	0.9033	0.1621	0.9005	0.9608	0.9767	5.6188	10.1702	1.0000	2.0000	4.6095
	ET	0.8785	0.1576	0.8708	0.9280	0.9567	4.0873	7.5946	1.0000	1.4142	3.0000
UNETR [6]	WT	0.9118	0.1020	0.9085	0.9460	0.9648	8.0959	14.7898	1.7321	3.1177	7.2801
	TC	0.8578	0.1989	0.8690	0.9312	0.9604	7.1861	12.2032	1.6168	3.0000	6.8367
	ET	0.8434	0.1863	0.8366	0.9036	0.9426	5.2969	9.0289	1.0000	1.7321	4.2202
MEU-Net (PA,4 Heads)	WT	0.9325	0.0671	0.9242	0.9557	0.9735	6.7218	12.1481	1.4142	2.4495	5.8732
	TC	0.9082	0.1486	0.9119	0.9632	0.9784	4.4011	7.3866	1.0000	1.7321	3.3535
	ET	0.8863	0.1452	0.8759	0.9279	0.9598	3.2808	6.0803	1.0000	1.4142	2.4495
MEU-Net (PA,3 Heads)	WT	0.9365	0.0658	0.9285	0.9575	0.9750	5.2830	8.0237	1.4142	2.2361	5.1962
	TC	0.9075	0.1553	0.9100	0.9630	0.9785	4.3778	7.8173	1.0000	1.4142	3.3166
	ET	0.8863	0.1475	0.8782	0.9325	0.9589	3.2174	6.4322	1.0000	1.4142	2.2361

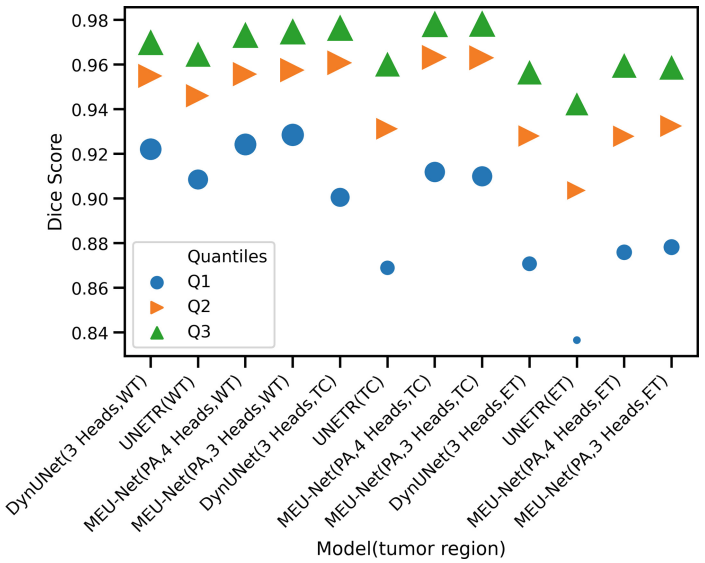

Fig. 3. Scatter plot comparing Dice Score quantiles of the models: MEU-Net (PA,4 Heads), MEU-Net (PA,4 Heads), DynUNet(3 Heads), and UNETR; on our internal validation-set

0.8445, and HD95 4.03, 10.16, 14.50. Ensemble-3 performs marginally better than ensemble-2 and is comparable to ensemble-1.

Table 3 shows the result (means, standard deviations, quartiles of Dice Scores and HD95) of the ensemble-3 assessed on the test set. Figure 5 compares the result of ensemble-3 assessed in the validation set and test set, and MEU-Net (PA, 4H) validated on the internal validation set. The ensemble-3 performs higher in the test set than the validation set for WT, TC, and ET.

Table 2. Validation-set results (219 subjects). *Ensemble-1*, *ensemble-2*, and *ensemble-3* are denoted by E-1, E-2, and E-1+E-2, respectively.

Ensemble models	Tumor regions	Dice Score					HD95				
		Mean	SD	Q1	Q2	Q3	Mean	SD	Q1	Q2	Q3
E-1	WT	0.9249	0.0749	0.9023	0.9434	0.9677	4.0403	6.4892	1.4142	2.2361	3.8708
	TC	0.8607	0.2102	0.8656	0.9411	0.9670	11.4049	50.1190	1.0000	1.7321	4.2500
	ET	0.8419	0.2194	0.8361	0.9002	0.9545	17.7092	73.6879	1.0000	1.4142	2.4495
E-2	WT	0.9210	0.0755	0.9005	0.9399	0.9669	4.6530	6.3841	1.7321	3.0000	4.8990
	TC	0.8588	0.2111	0.8572	0.9388	0.9663	9.9649	43.7501	1.0000	2.0000	4.2426
	ET	0.8364	0.2234	0.8325	0.9001	0.9531	17.7958	73.6651	1.0000	1.4142	2.4495
E-1+E-2	WT	0.9244	0.0748	0.9035	0.9433	0.9684	4.0345	6.4367	1.4142	2.2361	4.0000
	TC	0.8600	0.2078	0.8601	0.9385	0.9666	10.1635	44.0131	1.0000	2.0000	4.2426
	ET	0.8445	0.2100	0.8381	0.9012	0.9560	14.4990	65.3128	1.0000	1.4142	2.8284

Table 3. Test-set results (570 subjects). *Ensemble-3* is denoted by E-1+E-2

Ensemble models	Tumor regions	Dice Score					HD95				
		Mean	SD	Q1	Q2	Q3	Mean	SD	Q1	Q2	Q3
E-1+E-2	WT	0.9274	0.0930	0.9129	0.9562	0.9778	4.7680	17.0557	1.0000	1.7321	4.1108
	TC	0.8717	0.2456	0.9135	0.9607	0.9809	17.9765	71.5372	1.0000	1.4142	3.0000
	ET	0.8750	0.1798	0.8539	0.9314	0.9658	10.6618	55.5210	1.0000	1.0000	2.0000

5 Discussion

This work presents our modified U-Net architecture, denoted MEU-Net, incorporating multiple encoders and a shared decoder with attention blocks. We use it for the segmentation of three sub-regions of glioblastoma: whole tumor, tumor core, and enhancing tumor, from multi-parametric MRI scans. We trained multiple variations of the MEU-Net, as described in Sect. 3.4, on the 1251 training subjects of the BraTS challenge 2021. The models were evaluated on a separate validation set of 219 subjects provided by the organizers. Averaging the performance metrics across all tumor regions and all subjects, the three ensembles described above, *ensemble-1*, *ensemble-2*, *ensemble-3*, achieved Dice Scores of 0.8758, 0.8721, 0.8763, and HD95 distances of 11.05, 10.80, 9.56. *Ensemble-3* performs marginally better than *ensemble-2* and *ensemble-1* for the validation set. Also, *ensemble-3* performs better in the test set than the validation set. These findings signify that the ensembling of additional models increases robustness and generalizes well, as the ensemble model perhaps remains unaffected by the failure of a single model or a single independent CNN component [11]. The source code of experiments carried out, is available at https://github.com/MMIV-ML/brats2021.

A longer term aim is to evaluate a version of our segmentation pipeline integrated into the clinical workflow[1]. The goal is to provide a real-time clinically

[1] Using e.g. our research PACS setup at our local hospital region, https://mmiv.no/wiml/.

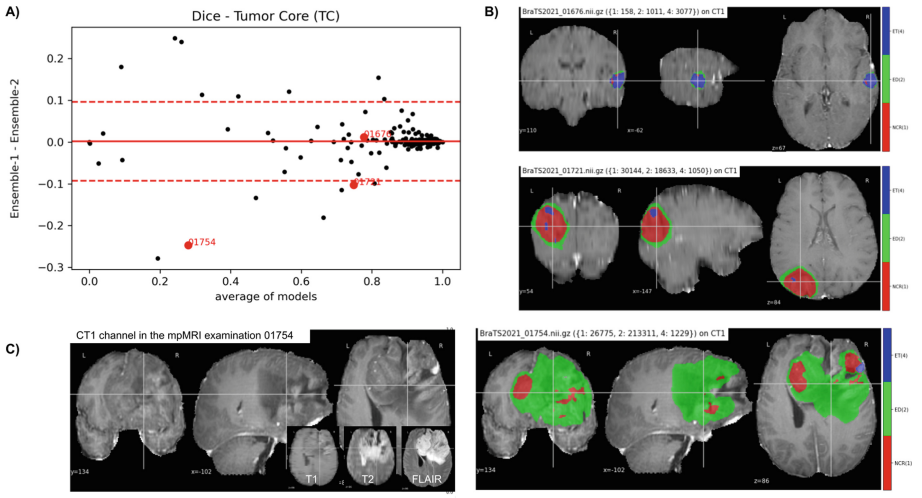

Fig. 4. A) Bland-Altman (BA) plot comparing Dice Scores for TC of *ensemble-1* and *ensemble-2* models, for the 219 subjects in the BraTS2021 validation dataset. The larger red dots with annotated ids refer to the two subjects in B) and the subject in C). **B**) Segmentation results from subjects 01676 and 01721 overlayed the mpMRI CT1 channel using the ensemble-1 model. **C**) Results from subject 01754. Here our model failed severely, with Dice Scores {MEU-Net segmentation vs. GT} of WT: 0.8320, TC: 0.1561, ET: 0.1194. Note the large and highly heterogeneous tumor mass seen in the CT1 channel without segmentation overlay, and the substantial variation in image quality among the four channels (small inserts depicts axial T1, T2 and FLAIR). Volumes of subregions NCR (red 1), ED (yellow 2) and ET (blue 4) are given in μL. The BA plot shows a difference in Dice Scores for TC between *ensemble-1* and *ensemble-2*, which prompts us to adopt *ensemble-3* using *ensemble-1&2*(as described in Sect. 3.4) aimed at achieving better generalization. (Color figure online)

relevant system that may be applied both in diagnostics and treatment planning. It can in principle be applied in the preliminary work up to define the tumor extent into the normal brain tissue and involvement of critical structures that might compromise resection. In the postoperative setting, the delineation of residual tumor is important for further therapy planning. However, the most interesting application is perhaps radiotherapy planning. The pipeline may be used for gross tumor volume delineation and contribute to more accurate delivery of irradiation. This may be particularly important for particle radiotherapy that allows limited irradiation to non-target tissue.

To be evaluated as a possible aid in established clinical workflows, it is especially important that the system is robust to naturally occurring variations in the input data. Currently, our model can completely fail in some cases (cf. Fig. 4C), with outputs that differ a lot from the "ground truth" segmentations produced by radiologists. Several factors may contribute to the problem, such as signal artifacts in MRI recordings, distribution shift, and that the models are trained

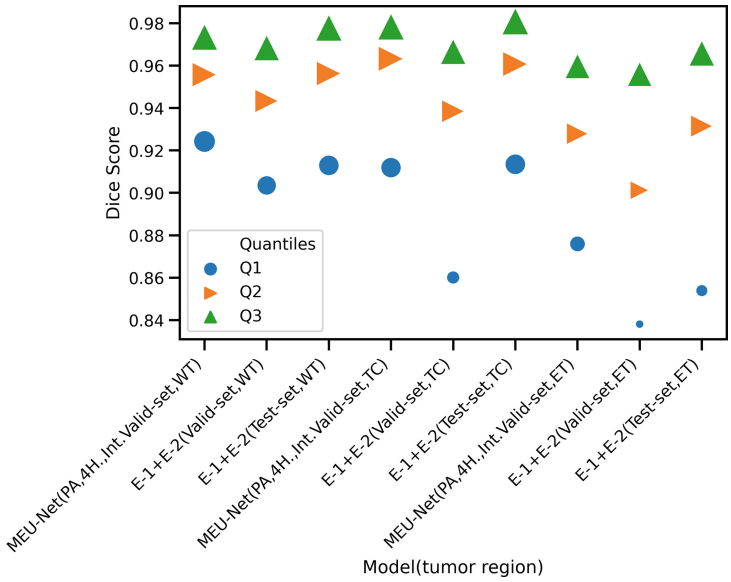

Fig. 5. Scatter plot comparing Dice Score quantiles of the models: MEU-Net (PA,4 Heads) on our internal validation-set(Int.Valid-set); E-1+E-2 on the organizer's validation-set(Valid-set); E-1+E-2 on the organizer's test-set. *Ensemble-1*, *ensemble-2*, and *ensemble-3* are denoted by E-1, E-2, and E-1+E-2, respectively. nH. denotes n deep supervision heads.

on limited data with too narrow diversity. We aim to tackle this by using various multi-pronged supervised and semi-supervised strategies, including more tailored augmentations and model architectures, and putting the human-in-the-loop. A crucial limitation of the current setup based on the BraTS challenge is the unavailability of clinical and radiological history for the BraTS subjects. Having all available MRI recordings for a given case available would enable sequential, longitudinal tumor analysis.

References

1. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv:2107.02314 [cs] (2021). http://arxiv.org/abs/2107.02314
2. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. Can. Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
3. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Can. Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
4. Bakas, S., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117

5. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
6. Hatamizadeh, A., Yang, D., Roth, H., Xu, D.: UNETR: Transformers for 3D medical image segmentation. arXiv:2103.10504 [cs, eess] (2021), http://arxiv.org/abs/2103.10504
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] (2015). http://arxiv.org/abs/1512.03385
8. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-Excitation Networks. arXiv:1709.01507 [cs] (2019). http://arxiv.org/abs/1709.01507. version: 4
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
10. Jiang, Zeyu, Ding, Changxing, Liu, Minfeng, Tao, Dacheng: Two-stage cascaded U-Net: 1st place solution to BraTS challenge 2019 segmentation task. In: Crimi, Alessandro, Bakas, Spyridon (eds.) BrainLes 2019. LNCS, vol. 11992, pp. 231–241. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46640-4_22
11. Kamnitsas, K., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. arXiv:1711.01468 [cs] (2017). http://arxiv.org/abs/1711.01468
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 (2017). https://doi.org/10.1016/j.media.2016.10.004, https://www.sciencedirect.com/science/article/pii/S1361841516301839
13. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik **29**(2), 102–127 (2019). https://doi.org/10.1016/j.zemedi.2018.11.002, https://www.sciencedirect.com/science/article/pii/S0939388918301181
14. Ma, N., Li, W., Brown, R., Wang, Y., et al.: Project-MONAI/MONAI: 0.6.0 (2021). https://doi.org/10.5281/zenodo.5083813, https://zenodo.org/record/5083813
15. McKinley, Richard, Meier, Raphael, Wiest, Roland: Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation. In: Crimi, Alessandro, Bakas, Spyridon, Kuijf, Hugo, Keyvan, Farahani, Reyes, Mauricio, van Walsum, Theo (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 456–465. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_40
16. Nyul, L., Udupa, J., Zhang, X.: New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging **19**(2), 143–150 (2000). https://doi.org/10.1109/42.836373, conference Name: IEEE Transactions on Medical Imaging
17. Oktay, O., et al.: Attention U-Net: Learning Where to Look for the Pancreas. arXiv:1804.03999 [cs] (2018). http://arxiv.org/abs/1804.03999
18. Pérez-García, F., Sparks, R., Ourselin, S.: TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Prog. Biomed. **208**, 106236 (2021). https://doi.org/10.1016/j.cmpb.2021.106236, https://www.sciencedirect.com/science/article/pii/S0169260721003102
19. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Hum. Brain Mapping **31**(5), 798–819 (2009). https://doi.org/10.1002/hbm.20906, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2915788/

20. Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas: U-Net: convolutional networks for biomedical image segmentation. In: Navab, Nassir, Hornegger, Joachim, Wells, William M.., Frangi, Alejandro F.. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Wright, L., Demeure, N.: Ranger21: a synergistic deep learning optimizer. arXiv:2106.13731 [cs] (2021), http://arxiv.org/abs/2106.13731