Optimizing Brain Tumor Segmentation: A Comparative Study of U-Net Architectures

Beatriz Moreira
Faculdade de Ciências
Universidade de Lisboa
fc54514@alunos.fc.ul.pt

Rute Patuleia
Faculdade de Ciências
Universidade de Lisboa
fc51780@alunos.fc.ul.pt

Tiago Assis
Faculdade de Ciências
Universidade de Lisboa
fc62609@alunos.fc.ul.pt

1. Introduction

Glioma is the most common tumor type to originate in the brain and arises from the glial cells. The World Health Organization (WHO) classifies brain tumors into grades 1 to 4 based on histologic features and molecular parameters, with grade 4 glioblastomas (GBM) being the most aggressive and lethal. Despite advancements in diagnosis and treatment, GBM mortality rates remain high, with patients typically surviving about 16 months with standard treatment. Extensive research to improve diagnosis, characterization, and treatment is paramount to reduce the mortality rate of this disease.

Glioma segmentation is crucial for tracking tumor growth, assessing treatment effectiveness, predicting survival, and planning treatment. Manual segmentation is used for this purpose, but is time-consuming, labor-intensive, and suffers from inter- and intra-examiner variability [12]. It involves identifying and outlining the extent of the whole glioma and its sub-regions on medical imaging scans, typically using magnetic resonance imaging (MRI).

The Brain Tumor Segmentation (BraTS) [1] Challenge is an annual competition organized by the Medical Image Computing and Computer-Assisted Interventions (MICCAI) aimed at identifying the most effective automatic segmentation algorithms for brain diffuse glioma patients, by freely providing fully annotated datasets every year to evaluate the latest state-of-the-art (SotA) approaches.

Convolutional neural networks (CNN) greatly dominate biomedical image segmentation, including brain tumor segmentation tasks [4]. More specifically, the encoder-decoder architectures with skip connections introduced by the U-Net have been outperforming most algorithms used in the BraTS challenge for the past 5 years and in the overall semantic segmentation field focused on medical imaging.

Our approach follows the top-ranked approaches which focus on variations of the U-Net architecture with added dilated convolutions, deep supervision, and BraTS-specific optimizations.

2. Problem Statement

The dataset used is the BraTS 2023 Challenge dataset (Synapse ID syn51156910) [2, 3, 8–10] which includes 1470 patients, consisting of 1251 cases annotated with the corresponding ground truth, used for training; and 219 cases with no annotation, used for the public leaderboard evaluation. Each case contains multi-parametric MRI scans with 4 modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). All the ground truth masks have been manually annotated and were approved by experienced neuro-radiologists. The image sizes of all scans and ground truth masks are $240 \times 240 \times 155$.

Mask annotations comprise the GD-enhancing tumor (ET, label 3), the peritumoral edematous/invaded tissue (ED, label 2), and the necrotic tumor core (NCR, label 1). The sub-regions considered in the challenge evaluation are the "enhancing tumor" (ET), the "tumor core" (TC), and the "whole tumor" (WT) (Figure 1). The TC entails both the ET as well as the NCR parts of the tumor. The WT describes the complete extent of the tumor, as it entails all the sub-regions. This creates the possibility of approaching this problem in two ways: consider it a multi-class classification problem and predict the label at each voxel; consider it as multiple binary classification problems by dividing the segmentation mask into three one-hot-encoded channels for each sub-region. The latter approach has been shown to improve performance on the BraTS challenge by optimizing each sub-region directly instead of the individual classes [13].

During this task, we expect to obtain as model outputs the multi-channel segmentations with the corresponding predicted tumor sub-regions delineated for each case. These results can be evaluated using the Dice coefficient and the 95% Hausdorff distance (HD) for each sub-region. The dice score compares the voxel-wise agreement between the prediction and the ground truth by computing the overlaps between the two, following Equation 1.

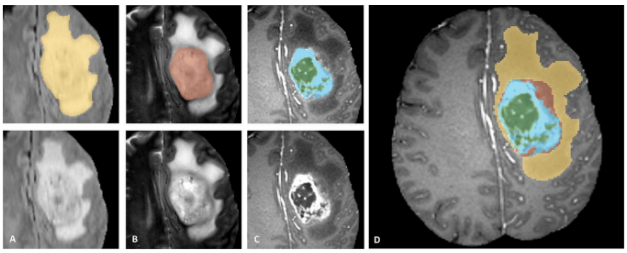

Figure 1. Manual annotation through expert raters. Image patches show from left to right: the whole tumor visible in FLAIR (A), the tumor core visible in T2 (B), the enhancing tumor structures visible in T1c (blue), surrounding the cystic/necrotic components of the core (green) (C). The segmented sub-regions are combined to generate the final labels of the tumor structures (D). Adapted from [9].

$$\text{Dice coefficient} = \frac{2|P \cap T|}{|P| + |T|} \tag{1}$$

The HD computes the maximum distance of all the distances between the boundary points of the prediction to their nearest boundary point in the ground truth. It measures how far the most mismatched point in one set is from any point in the other set (Equation 2). The 95% Hausdorff distance (HD95) is a variation of this calculation where only the 95^{th} percentile of the distances are taken into account.

$$\text{HD}(P, T) = \max_{p \in P} \min_{t \in T} \|p - t\|_2 \tag{2}$$

3. Technical Approach and Methodology

Our work implemented previous top-ranked approaches for the BraTS challenge, which focus on improvements to the U-Net architecture, and challenge-specific data processing pipelines and optimizations. We followed the approaches of Henry et al. [5] and Isensee et al. [7], by implementing their network architecture designs.

The data preprocessing pipeline starts by loading the MRI scans provided as NIfTI files, a format widely used by neuroimaging software. By trial and error, images were then cropped to a fixed size that minimized the amount of background voxels to obtain the smallest bounding box containing the whole brain. This helps the network to not focus on classifying voxels as background and learn from an almost sole view of the brain. Cropping was not applied to the test set images. Since the MRI intensities in each case vary depending on both the hardware and software used, a standardization step was performed for each scan by subtracting the mean and dividing by the standard deviation considering only the brain regions (non-brain/background voxels remained at 0). Finally, the four scans were stacked and downsampled to a fixed size, resulting in tensors of shape $4 \times 128 \times 128 \times 128$.

The ground truth segmentations were equally loaded and cropped. The provided labels (ED, ET, NCR) present in each segmentation are not directly the ones used for the evaluation of the segmentations. Thus, we take the approach to preprocess each mask taking into account the three partially overlapping sub-regions (ET, TC, WT) used for evaluation. This was done by stacking three one-hot-encoded channels corresponding to the three sub-regions: for the ET channel, the voxels of interest are those identified as ET in the original ground truth; for the TC channel, those identified as either ET or NCR; and for the WT channel, those identified as either ED, ET, or NCR. The segmentations were finally downsampled to a tensor shape of $3 \times 128 \times 128 \times 128$.

Data augmentation can be used to enlarge the training set and improve generalization artificially. Data augmentation is applied on the fly during training and follows the steps in [6], consisting of random rotations, scaling, gaussian noise, gaussian blur, brightness, contrast, simulated low resolution, gamma, and mirroring, that apply to a volume with a certain probability each. If an augmentation occurs, the corresponding ground truth suffers the same augmentations only if they are rotations, scaling, and mirroring. Augmentations were only applied to the training set.

The first architecture implemented [5] has a U-Net like design, consisting of four symmetric downsampling (encoder) and upsampling (decoder) stages, with skip connections from the encoder to the decoder at each step, where concatenation of information occurs. Each stage consists of two $3 \times 3 \times 3$ convolutions with stride 1 and 'same' padding, and a MaxPool layer with a kernel size of $2 \times 2 \times 2$ and stride 2. The initial number of filters is 48 and it is doubled after each stage, while the spatial resolution halves. After the last stage, two $3 \times 3 \times 3$ dilated convolutions with dilation 2 were performed, and the output of the last encoder stage was concatenated to the output obtained after these dilated convolutions. A final convolution is performed to restore the number of filters to the same value before the dilated convolutions.

The decoder is an almost symmetric copy of the encoder, where spatial upsampling is performed using trilinear interpolation. The last convolution layer used a $1 \times 1 \times 1$ kernel with 3 output channels and a sigmoid activation.

According to the authors of the paper, the dilation trick is used to perform a pseudo fifth stage without further downsampling of the feature maps, reducing the possibility of irreversible information loss due to the small input resolution size. For this network, instance normalization and ReLU activation functions followed each convolution.

The second network implemented [7] also has a U-Net like design, similar to the previous architecture, but consisting of 5 downsampling and upsampling stages, with a bottleneck between them. The initial number of filters was 32.

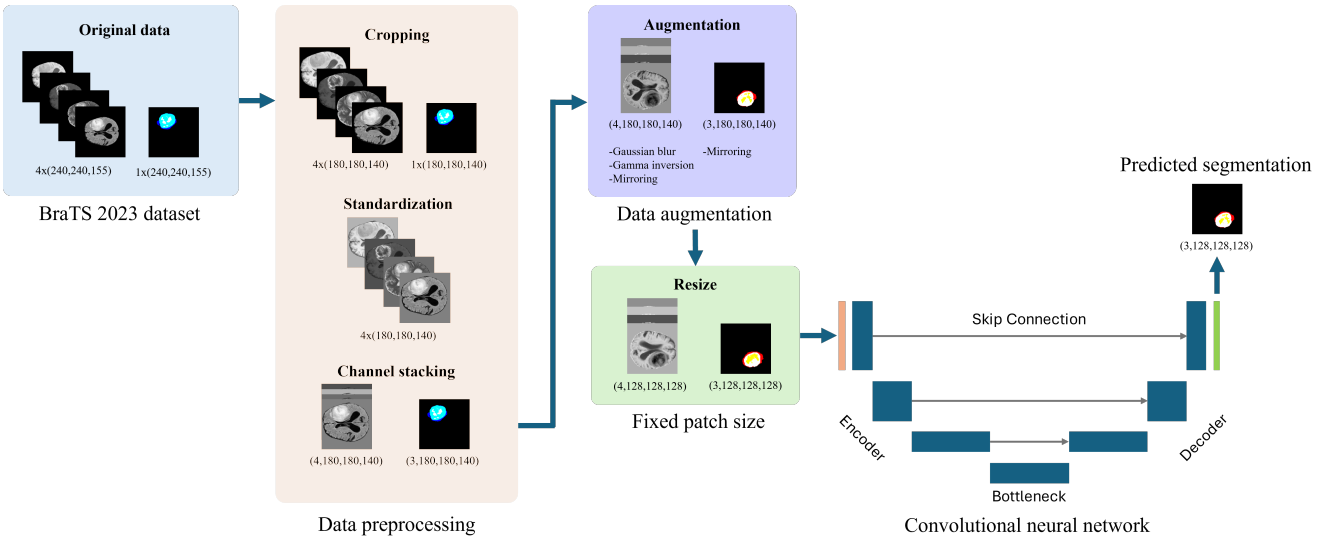

Figure 2. Methodology pipeline used, including data preprocessing, augmentation, downsampling, and modeling through U-Net-like networks.

In the bottleneck, double convolutions occur, but the number of filters is maintained. The upsampling in the decoder is performed using transposed convolutions with kernel size $2 \times 2 \times 2$ and stride 2. Each convolution layer is followed by instance normalization, and Leaky ReLUs (10^{-2} negative slope) are used as non-linear activation functions. Both of these architectures produce several auxiliary outputs used for deep supervision in different stages of the decoder (Figure 2).

As detailed in section 3, we implemented our versions of previous solutions to the BraTS Challenge. The experiments focused on assessing the baseline segmentation metric scores and generalization capabilities. The contribution of specific components of the networks to the overall model performances by performing ablation studies. We trained both the Dilated U-Net and nnU-Net architectures from scratch on the same 80% training and 20% validation set splits and evaluated their performance in terms of dice scores and HD95 to look for region-specific differences between the approaches.

The baseline for Dilated U-Net included an Adam optimizer with a starting learning rate of 10^{-4}, and a cosine annealing scheduler to decay the learning rate over time. The baseline for nnU-Net included an SGD optimizer with a learning rate of 10^{-2} and Nesterov momentum of 0.99, and a polynomial scheduler with an exponent of 0.9. Both networks on default configuration were trained using the dice loss, which is the complementary of the dice score, computed for all samples in the batch at once, and equal to the mean value of the dice losses per channel. Additionally, nnU-Net used the binary cross-entropy loss as a second loss function, thus its total loss during training was the sum of both the dice and cross-entropy losses. Dilated U-Net used the dice loss as the sole loss function. Both networks used a batch size of 2.

The ablation studies investigated the removal of deep supervision, and changing optimizers, schedulers, and loss functions, as well as the addition of weight decaying. The combinations tested and their performance metrics are presented in Table 1. All combinations employed the same starting learning rate of 10^{-3}.

All experiments were conducted on a single Nvidia Tesla P100 GPU with 16 GB VRAM in Kaggle kernels.

4. Results

Preliminary tests were performed with baseline architectures that were trained for 100 epochs (Figure 3). Looking at the plots more deeply, both show a rapid increase in their Mean Dice Score initially, and then a plateau, suggesting that the models learn very effectively early, and stabilize later on. However, this could also be a direct consequence of the relatively small dataset used, meaning that the models quickly learn the patterns of the available data and then plateau. This may also suggest that the models may have not been deep or complex enough to fully capture the complexities of the dataset. Another additional constraint may be the parameters used in this process, such as the learning rate. Regarding the first reason, this problem can be reduced by performing more aggressive data augmentations or by creating synthetic data from the original dataset, as performed by the winning solution for BraTS 2023 [11]. For the second reason, more complex models should be employed, as the models implemented in this work derive from the BraTS

Table 1. Ablation studies' evaluation results with the testing set.

Model	Mean Dice	Dice ET	Dice TC	Dice WT	Mean HD	HD ET	HD TC	HD WT	Sens	Spec
Dilated U-Net										
Baseline	.7448	.6320	.7556	**.8469**	16.98	30.84	**12.48**	7.62	**.8283**	.9987
- S	.6988	.5924	.6951	.8088	23.31	37.38	18.38	14.18	.7625	.9983
+ SGD	.7504	.6539	.7520	.8452	17.22	28.18	14.00	9.48	.7677	**.9992**
+ P	.7037	.5981	.7197	.7934	24.62	38.49	17.62	17.76	.7762	.9985
+ 2L	**.7536**	**.6525**	**.7614**	**.8468**	**15.80**	**25.96**	14.16	**7.29**	.7703	**.9991**
+ WD + 2L	.7468	.6364	.7586	.8455	17.69	31.89	13.56	7.61	.7781	.9990
nnU-Net										
Baseline	0.7314	0.6443	0.7227	0.8273	20.11	28.51	20.84	10.98	0.7667	0.9989
- S	0.7392	0.6428	0.7354	0.8393	18.39	29.37	16.40	9.41	0.7547	0.9991
+ CA	0.7480	0.6507	0.7453	**0.8480**	15.76	23.45	16.88	6.94	0.7644	0.9990
+ A	0.7488	0.6325	**0.7663**	0.8476	17.60	33.75	**11.43**	7.61	**0.8068**	0.9989
+ A + WD	**0.7513**	**0.6514**	0.7577	0.8447	**15.21**	**24.89**	13.16	**7.59**	0.7473	**0.9993**
+ A + CA + 1L	0.7077	0.6028	0.7077	0.8125	22.58	32.55	19.72	15.48	0.7790	0.9984

S - Deep Supervision CA - Cosine Annealing Scheduler P - Polynomial Scheduler A - Adam Optimizer SGD - Stochastic Gradient Descent Optimizer
WD - Weight Decay 1L - 1 Loss Function 2L - 2 Loss Functions Sens - Sensitivity Spec - Specificity

2020 challenge, which comprises an even smaller dataset that might not require very complex models. Finally, lowering the learning rate would require longer training times, which we were limited by in this project. In both cases, the training and validation curves align, which suggests good generalization and no signs of overfitting.

Although the plateau happens around epoch 40, all further results were obtained by training both networks for 20 epochs in order to balance training time and performance, as every 15 epochs took 10 hours to complete with our resources.

The models showed good segmentation accuracy, as shown in the examples presented in Figure 4 (mean Dice scores of 0.8676 and 0.8807, respectively). The different tumor sub-regions were mostly delineated with a high score. The hardest sub-region to delineate was ET, while WT was easily identified. Overall, both architectures achieved high segmentation scores, however, Dilated U-Net generally outperformed nnU-Net in terms of mean scores. This suggests that the Dilated U-Net architecture, with its dilated convolutions, might be more effective in capturing intricate details within the tumor regions.

For the ablation studies, we employed different optimization techniques, including variations of optimizers (Adam, SGD), learning rate schedulers (Cosine Annealing, Polynomial), and weight decay. The optimization strategies had a relative impact on segmentation performance, with certain configurations yielding better results than others for each architecture. For example, nnU-Net models trained with the Adam optimizer and Cosine Annealing scheduler tended to perform better compared to those trained with other configurations and the baseline (which uses SGD and a polynomial scheduler). Additionally, the addition of weight decay seems to improve this network's results. For Dilated U-Net, the use of 2 loss functions (instead of solely one) improved the performance results substantially, on the contrary, removing deep supervision yielded a drastic drop in performance.

5. Discussion

In the present study, we explored variations of the U-Net architecture for the segmentation of glioma tumors, with a focus on improving segmentation scores and exploring how several components affect the overall model performances. The results of our implementations, including Dilated U-Net and nnU-Net, underscore the potential of these architectures in accurately segmenting different tumor sub-regions, namely, whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The effectiveness of these models in delineating these sub-regions was assessed using Dice coefficient and 95% Hausdorff distance (HD95).

Our findings showed that both Dilated U-Net and nnU-Net architectures achieved high segmentation accuracy. However, the Dilated U-Net obtained better segmentations, which might be explained by the fact that this architecture uses dilated convolutions, which expand the receptive field without further losing spatial resolution in downsamplings. This architecture is then more likely to capture complex details within tumor regions, leading to improved segmentation results.

Among the three sub-regions, the ET posed the great-

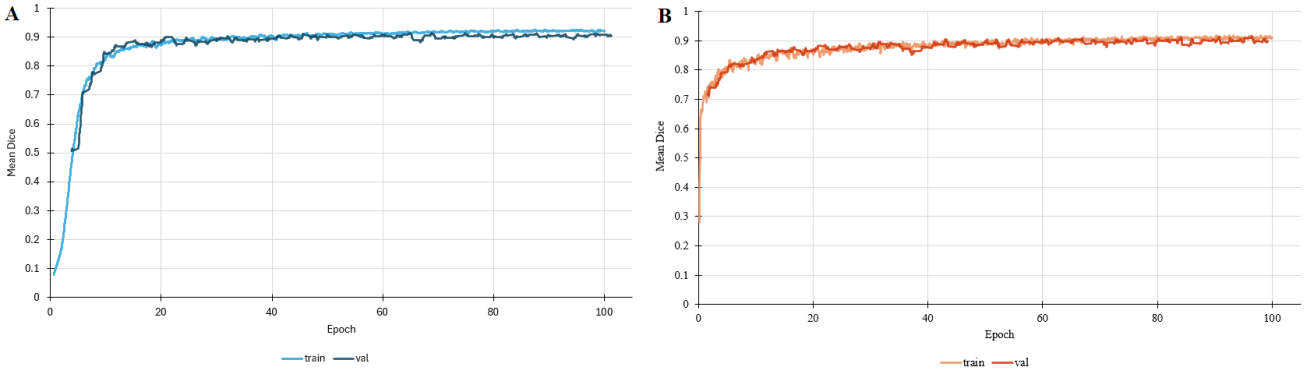

Figure 3. Training and validation mean Dice scores. (A) Dilated U-Net. (B) nnU-Net.

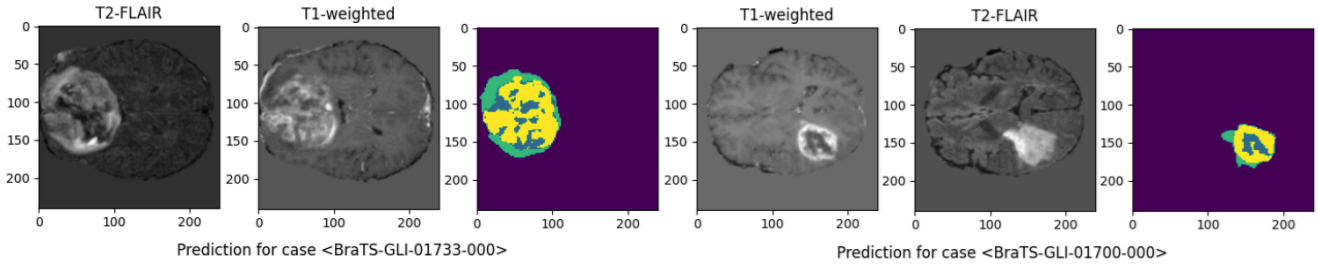

Figure 4. Slices of the obtained segmentations for two case examples in the BraTS 2023 challenge dataset using the Dilated U-Net. Mean Dice scores of 0.8676 and 0.8807, respectively

est challenge for segmentation. The ET region exhibited high variability in intensity and shape across different patients, having very irregular patterns, making it harder for the model to segment it consistently.

Furthermore, the effectiveness of our models was improved by incorporating data augmentation and deep supervision techniques. Data augmentation techniques, encompassing a range of transformations (random rotations, scaling, gaussian noise, gaussian blur, brightness, contrast, simulated low resolution, gamma, and mirroring) applied during training, enhanced the robustness of the models to variations in input data, contributing to improved generalization. Deep supervision, involving the introduction of auxiliary outputs at different stages of the network, facilitated more consistent performance across network layers, leading to more accurate and reliable segmentations. However, despite the promising results obtained, our study has some limitations. The relatively small dataset used in this implementation may have constrained the models' ability to capture the complexities of glioma tumor characteristics fully. Additionally, the chosen optimization strategies and hyperparameters may not have been fully optimized, leaving room for further exploration and refinement. Thus, future work should involve using more sophisticated feature extraction and learning approaches or mechanisms to better predict the variability of the ET region.

In conclusion, the present study showed the efficacy of Dilated U-Net and nnU-Net architectures in glioma tumor segmentation, highlighting the importance of architectural design choices, optimization strategies, and the incorporation of advanced techniques such as data augmentation and deep supervision. These findings contribute to advancing the field of medical image segmentation and hold promise for improving diagnosis, treatment planning, and patient outcomes.

References

[1] BraTS. Challenge. https://www.synapse.org/#!Synapse:syn53708126/wiki/626320, . Accessed: 04/05/2024. 1

[2] BraTS. BraTS 2023. https://www.synapse.org/#!Synapse:syn51156910/wiki/621282, . Accessed: 04/05/2024. 1

[3] Bakas S. et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Nature Scientific Data*, 4, 2017. https://doi.org/10.1038/sdata.2017.117. 1

[4] Fernando K. et al. Deep and statistical learning in biomedical imaging: State of the art in 3d mri brain tumor segmentation. 2023. https://doi.org/10.1016/j.inffus.2022.12.013. 1

[5] Henry T. et al. Top 10 brats 2020 challenge solution: Brain tumor segmentation with self-ensembled, deeply-supervised 3d-unet like neural networks. 2020. https://doi.org/10.48550/arXiv.2011.01045. 2

[6] Isensee F. et al. Automated design of deep learning methods for biomedical image segmentation. *Nature Methods*, 2020. https://doi.org/10.1038/s41592-020-01008-z. 2

[7] Isensee F. et al. nnu-net for brain tumor segmentation. 2020. https://doi.org/10.48550/arXiv.2011.00848. 2

[8] Karargyris A. et al. Federated benchmarking of medical artificial intelligence with medperf. *Nature Machine Intelligence*, 5:799–810, 2023. https://doi.org/10.1038/s42256-023-00652-2. 1

[9] Menze B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. https://doi.org/10.1109/TMI.2014.2377694. 2

[10] Ujjwal B. et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. 2021. https://doi.org/10.48550/arXiv.2107.02314. 1

[11] André Ferreira, Naida Solak, Jianning Li, Philipp Dammann, Jens Kleesiek, Victor Alves, and Jan Egger. How we won brats 2023 adult glioma challenge? just faking it! enhanced synthetic data augmentation and model ensemble for brain tumour segmentation, 2024. https://doi.org/10.48550/arXiv.2402.17317. 3

[12] Diana Veiga-Canuto, Leonor Cerdá Alberich, Cinta Nebot, Blanca Heras, Ulrike Pötschger, Michela Gabelloni, Jose Carot, Sabine Taschner-Mandl, Vanessa Düster, Adela Cañete, Ruth Ladenstein, Emanuele Neri, and Luis Marti-Bonmati. Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images. *Cancers*, 14:3648, 2022. https://doi.org/10.3390/cancers14153648. 1

[13] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. *Automatic Brain Tumor Segmentation Using Cascaded Anisotropic Convolutional Neural Networks*, page 178–190. Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-75238-9_16. 1