Neglectable effect of brain MRI data preprocessing for tumor segmentation.

Ekaterina Kondrateva\textsuperscript{1,2}, Polina Druzhinina\textsuperscript{1,2}, Alexandra Dalechina\textsuperscript{3}, Boris Shirokikh\textsuperscript{1}, Mikhail Belyaev\textsuperscript{1,2}, and Anvar Kurmukov\textsuperscript{2}

\textsuperscript{1} Skolkovo Institute of Science and Technology, Russia, Moscow
\textsuperscript{2} Artificial Intelligence Research Institute, Russia, Moscow
\textsuperscript{3} National Medical Research Center for Neurosurgery, Moscow Gamma Knife Center
\textsuperscript{4} ekaterina.kondrateva@skoltech.ru

\textbf{Abstract.} Magnetic resonance imaging (MRI) data is heterogeneous due to the differences in device manufacturers, scanning protocols, and inter-subject variability. A conventional way to mitigate MR image heterogeneity is to apply preprocessing transformations, such as anatomy alignment, voxel resampling, signal intensity equalization, image denoising, and localization of regions of interest (ROI). Although preprocessing pipeline standardizes image appearance, its influence on the quality of image segmentation and other downstream tasks on deep neural networks (DNN) has never been rigorously studied. Here we report a comprehensive study of multimodal MRI brain cancer image segmentation on TCIA-GBM open-source dataset. Our results demonstrate that most popular standardization steps add no value to artificial neural network performance; moreover, preprocessing can hamper model performance. We suggest that image intensity normalization approaches do not contribute to model accuracy because of the reduction of signal variance with image standardization. Finally, we show the contribution of skull-stripping in data preprocessing is almost negligible if measured in terms of clinically relevant metrics. We show that the only essential transformation for accurate analysis is the unification of voxel spacing across the dataset. In contrast, anatomy alignment in form of non-rigid atlas registration is not necessary and most intensity equalization steps do not improve model productiveness. We propose to re-establish brain segmentation preprocessing pipelines, removing all redundant steps, which in turn, drastically improves the reproducibility of the results.

\textbf{Keywords:} brain MRI · preprocessing · nn-Unet.

1 Introduction

In recent years, modern deep neural networks (DNN) have steadily improved the quality of automatic segmentation pipelines in medical imaging. Specifically, for the task of brain tumor segmentation, the performance of DNNs has achieved human-level efficiency \cite{5}. This advancement can be explained by improving
DNNs architectures (training pipelines) and growth of the training datasets. For example, the size of the BraTS competition [2] training dataset increased from one hundred subjects in 2013 to two thousands subjects in 2021. Simultaneously, top-performing algorithms progress from random forest and gradient boosting trees on radiomics features, first, to fully convolutional networks; next, to u-shaped UNet and UNet-like networks; and finally, to Vision transformers.

In contrast, the preprocessing steps used to prepare data for analysis seem to have undergone considerably fewer changes. For instance, the set of preprocessing steps for brain MRI images has remained relatively stable and has been reproduced across the majority of papers on the topic from early 2010s till nowadays. Here we challenge the conventional pipelines for MRI image processing and question their necessity for accurate prediction with regard to new advanced deep learning machinery.

The traditional brain MRI preparation steps could be divided into four distinct categories. First, is subject-wise image alignment, typically in a form of rigid registration of one MRI sequence to another, (e.g. T2-FLAIR onto T1 with contrast enhancement). This step is mandatory if one uses multiple MR modalities to predict a single segmentation map or training 3D model, and to ensure correct alignment between ground truth annotation and corresponding image’s voxels. The second category is voxels resampling to some standard. The most common methods are non-rigid registration to some anatomical atlas and voxel resampling to homogeneous spacing (often 1mm$^3$). The third includes steps which affect voxels’ intensity distribution, such as bias-field correction, intensity normalization (typically in a form of image-wise z-scoring), image denoising methods (e.g., SUSAN [24]), and histogram equalization [17]. Finally, the last step, that is preserved in almost all the papers, is skull stripping as a method to localize regions of interest (the brain tissue) or implement feature selection to ease localization and reduce the amount of False Positives.

While the motivation behind applying these steps is clear: to standardize image appearance and remove different sources of domain shift [13], they are computationally costly and their utility for image segmentation lacks investigation. It is widely known that increasing variability of the data by data augmentation (image resizing, non linear intensity transformations, applying noise, etc.) leads to the improved DNNs performance [20]. However, data preprocessing works quite in the opposite way by reducing data variance. From this perspective, it might be the right time to reconsider the necessity of complete data standardization.

In this study we analyze most popular preprocessing steps for brain MRI preprocessing pipeline and measure their influence on tumor segmentation task. We analyze different preprocessing strategies and recommend the minimal pipeline required for accurate segmentation with the benefits of lower computational costs.
2 Related works

Medical images’ preprocessing is a de-facto standard first step in almost all deep learning pipelines for medical image segmentation. For some modalities, such as computed tomography (CT), it is less mandatory, because CT voxels value corresponds to tissue attenuation of radiography signal. But for MRI data most of the studies use at least some preprocessing. In the present work, we analyze the influence of the seven most popular MRI preprocessing steps on the task of brain tumor segmentation.

The goal of MRI preprocessing is to standardize images’ appearance. Each step targets its’ own source of domain shift. Respectively, histogram equalization targets the differences in voxels’ distribution by means of non-linear transformation, a simpler monotonic analog is a linear transformation by shift and scaling, either in a form of z-scoring (zero mean and unit variance) or by min-max scaling (making minimal voxel value to be 0 and maximal to be 1); denoising algorithms target the level of noise [24]; bias-field correction, mitigates the effects of magnetic field variation[5]; voxel resampling interpolates images of lower dimensions. The latter along with intra-subject rigid registration, target inter-modality differences (depending on clinical needs some MR sequences, within a single study, might be acquired in higher resolution). Skull stripping is performed to localize the region of interest and filter out False Positive voxels. To attempt to standardize the anatomic structure of images within the data collection, researchers use non-rigid image registration to some anatomic atlas.

Interestingly, while the general scope of MRI preprocessing methods is set, there is no consensus regarding which of them should be used, in what configuration, and in which order. For instance, all observed papers analyzing multi-sequence MRI use inter-modality registration [9, 11, 13, 14, 18, 21]; most of the papers use some kind of voxel resampling, either to an isotropic voxel (e.g. 1×1×1mm$^3$), or to the same image resolution (in voxels), or both, by means of non-rigid atlas registration [9, 11, 13, 14, 18, 21, 12]. Two exceptions are [12], where authors train on retrospective data (with inclusion/exclusion criteria) and test on prospective data, collected with unified scanning protocol; and [25], who again, used data acquired with the same scanning protocol thus, almost no domain variability. In contrast, multi-institutional studies typically use some intensity or noise correction approaches, with most popular being: histogram matching [9, 11, 18], bias-field correction [9, 11, 12], denoising [11, 12] and z-scoring [21, 11, 13, 14, 22]. Finally, all but one observed paper [25] use skull stripping, arguing that non-brain tissue is a significant source of error for downstream tumor segmentation [5].

In general, researchers experimenting on single-institutional data, or data collected under unified acquisition protocol tend to use fewer preprocessing steps. On the contrary, the analysis of heterogeneous multi-center data typically in-
cludes more standardization. In present work, we demonstrate that even for relatively small data collection, the effect of most standardization steps is negligible both for mathematical metrics, such as Sorensen-Dice coefficient, Precision, and Recall, and clinically relevant characteristics, such as volumetric measurements.

To the best of our knowledge, by far there were two similar attempts to carefully analyze the influence of preprocessing steps for medical imaging tasks. Authors of \cite{15} test how preprocessing influence radiomics features calculation, and \cite{20} investigate the influence of preprocessing for three medical imaging tasks.

3 Methods

3.1 Data description

We report our results on a publicly available MRI Glioblastoma Multiforme (GBM) dataset\cite{23,6} from The Cancer Imaging Archive\cite{4}. The sample was chosen among other open-source collections by the following criteria: original DICOM images, with manual or semi-automatic lesion segmentation in native space, multimodal input for subjects, image heterogeneity as well as the overall sample size.

DICOM-SEG TCIA GBM\cite{4} dataset includes 102 unique patient. For each patient, the minimal set of four image modalities is available: T1-weighted (T1), T1 contrast-enhanced (CT1), T2-weighted (T2) and FLAIR images (FLAIR), and corresponding semi-automatic segmentation (textttGLISTRBoost\cite{3} with manual correction). The collection is comprised out of data from more than four manufacturers and 39 different study protocols. To convert provided DICOMs to Nifty format we use dicom2niix\cite{8}.

3.2 Experimental design

For all experiments, we start with rigid registration (no image resizing, only rotation and shift) of every MR sequence on CT1 as implemented in \cite{1}. Before model training (after all other preprocessing steps) we use image-wise z-scoring: \( X_s = \frac{X - \text{mean}(X)}{\text{std}(X)} \). Our goal is to test how different preprocessing steps affect segmentation quality. Here we provide a list of these steps along with implementation details.

Image size standardization. To align different images within the subject and different samples (all four MR sequences) between the subjects, we compare three resampling approaches. First, is resizing image to the same voxel size of \([240, 240, 155]\), second, resampling voxels’ size to isotropic \([1, 1, 1]\) mm\(^3\) resolution, and finally both, by means of non-rigid atlas (SRI24) registration as implemented in \cite{1}.

\footnote{https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=41517733}

\footnote{https://www.nitrc.org/plugins/mwiki/index.php/dcm2nii:MainPage}
Voxels’ intensity standardization. We have used three algorithms of images intensity correction: bias-field correction, denoising and histogram standardization. For bias-field correction we use N4 algorithm, implemented in Simple-ITK as a containerized solution from CaPTk toolbox. MR image nonlinear smoothing and denoising was performed with SUSAN [24] algorithm, as implemented in FSL. These steps were applied for each image individually, after the rigid registration. To standardize voxels’ distribution between the subjects we apply histogram equalization [17], as implemented in torchio library [19]. We estimate the parameters of histogram matching on training folds and apply the same transformation on the data from test folds. Histogram matching was applied modality-wise, meaning that we equalize voxels’ histograms for each MR sequence separately.

Skull stripping. Skull stripping is a method to localize the region of interest and filter out potential False Positives. While this is a conventional approach, which is almost always used, it is highly sensitive to algorithm choice, increases inference time and complicates pipeline reproducibility. To test how skull stripping affects model performance we use the best online available tool [11] HD-BET, which is an ensemble of multiple DNNs. We found that skull stripping can also occasionally strip labeled brain regions, and individual masks tend to cut off more information then shared ones (see Appendix B). Therefore, we decided to estimate brain mask on CT1 and apply it to all other MR modalities after image alignment.

3.3 Model architecture and training

In the study we focused on 3D segmentation and use implementation of nn-UNet [10] model as a standard benchmark solution in medical segmentation.

We train 3D nn-UNet model with multi-modal input of four modalities with the following parameters: patch size of \([128,128,128]\) voxels with batch size of two. For the ablation study we train using combined DiceCELoss loss and learning rate 0.0008 with weight decay 0.0001; and Adam optimizer with momentum 0.99. These are default learning parameters for the similar 3D segmentation task on BraTS2021 [9].

All ablation experiments were trained for 100 epochs without data augmentations, as data augmentation can interfere with measurements of preprocessing effect. Two experiments for more detailed comparison were trained till the convergence with data augmentations, included zoom, flips, noise, blur, brightness, contrast and implemented in NVIDIA/DALI [10] library. While training with data augmentations we follow 20 epochs patience as stopping criterion.

All experiments were performed on a 3-fold cross-validation setup. The average time for training and inference in three folds for one experiment was 12 hours on 32GB NVIDIA Tesla V100 PCIe.

9 github.com/NVIDIA/DeepLearningExamples/tree/ddbcd54056e8d1bc1d5d556b6811d12b4e8ab34c570beea1947/PyTorch/Segmentation/nnUNet
10 https://github.com/NVIDIA/DALI
All the code for training and preprocessing is available in the repository\textsuperscript{11}. We use \textit{Dice}, \textit{Surface Dice} and \textit{Hausdorff\textsuperscript{95}} metrics, along with \textit{Sensitivity}, \textit{Specificity} and \textit{Precision} to assess the segmentation quality. All calculations were performed in the original CT1 image space. To measure the error in tumor volume estimation we use \textit{Mean Absolute Error}, which is the simplest clinically relevant feature\cite{5}. To measure the differences between voxels' intensities of healthy and tumored tissues we use \textit{Kullback-Leibler} divergence between corresponding intensities histograms.

The significance of difference were assessed by \textit{Wilcoxon} ranked-sign test \textit{Bonferroni} corrected for multiple comparisons. All metrics measurements were averaged over all subjects on a test fold.

4 Results

4.1 Preprocessing ablation study

Our first result is that at least one method of voxel resampling is mandatory to achieve voxels' alignment between different images of the same subject. All segmentation results (WT, TC, ET) without resampling are statistically significantly worse compared to the experiments with resampling, see Table 1 versus 2,3,4 (42.9, compared to 85.3, 85.9, and 86.4 for Whole Tumor (WT)). Second, we compare three ways to interpolate images, using either voxel resampling, interpolation to the same size, or both by means of anatomic atlas registration. We do not have a decisive winner, as each approach outperforms others exactly for one class: atlas registration is better for WT; voxels' resampling for tumor core (TC) and image resizing for the enhancing tumor (ET). Even though all differences are statistically significant, all p-values are less than 0.001 after the correction. We argue that these differences are negligible, and one should use the fastest approach. To make the final decision we further train 3. Resample to spacing and 4. Atlas registration models with the addition of extensive data augmentations, described in section 3.3. Models performance revealed to be equal on WT segmentation: 86.8 for voxel resampling versus 86.9 for atlas registration, voxel resampling is statistically significantly better for two remaining classes (Dice scores 62.6 and 68.5 for TC, and 57.9 and 60.6 for ET). We conclude that in real-life scenarios (more training epochs, extensive data augmentation) voxel resampling results in better performance and is easy to implement. Third, we compare experiments with different intensity normalization steps (Table 1 4a-c) to experiment only with atlas registration (Table 1 4). In terms of the Dice score b-f correction shows worse results for all classes compared to atlas registration, except for TC. Denoising, also shows inferior segmentation results in terms of Dice compared to the baseline, except for the WT class. Finally, histogram matching is worse in all comparisons (84.8 Dice vs 86.4 DICE in experiment 4, for WT). We argue that all intensity normalization steps could be omitted.

\textsuperscript{11} *
without hampering the performance. Finally, we test how brain extraction influences performance. We observe that for the WT segmentation skull stripping application gives a statistically significant jump in performance in terms of both Dice score \( (87.6, \text{p-value } 0.003) \) and Precision\( (88.1, \text{p-value } 0.002) \). However, for both other classes, it either decreases the performance (TC) or is statistically indistinguishable (ET). In addition, for volumetric measurements of WT the differences of the model with and without skull stripping are not statistically significant (MAE 9.2 ml for skull stripping and MAE 10.3 ml without skull stripping, p-value 0.112), thus from a clinical perspective, this additional step is not completely justified. Complete volumetric measurements for all experiments are provided in the Appendix A.

### Table 1. nn-UNet segmentation performance for one class on training for 100 epochs, comparison on GBM dataset: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). Statistically significant difference marked with green. Experiments \([3,4]\) compared to 2, \([4a-e]\) compared to 4. Dice and Precision are multiplied by 100.

|                      | WT       | WT       | ET       | ET       | TC       | TC       | Time, s |
|----------------------|----------|----------|----------|----------|----------|----------|---------|
| 1. Registration      | 42.9     | 76.0     | 36.8     | 65.1     | 30.5     | 50.6     | 6       |
| 2. Resample to spacing | 85.3   | 85.6     | 67.2     | 60.2     | 61.7     | 77.4     | 2       |
| 3. Resample to size   | 85.9     | 87.7     | 64.3     | 55.5     | 62.7     | 76.9     | 2       |
| 4. Atlas registration | 86.4     | 86.6     | 62.1     | 52.8     | 59.6     | 79.6     | 4       |
| 4a. Bias-field correction | 84.9   | 86.0     | 63.1     | 55.2     | 58.5     | 78.0     | 60      |
| 4b. Denoising         | 86.5     | 87.3     | 62.0     | 52.2     | 61.3     | 77.8     | 160     |
| 4c. Histogram matching | 84.8     | 86.8     | 58.6     | 48.4     | 57.6     | 76.6     | 6       |
| 4d. Skull stripping   | 87.6     | 88.1     | 61.3     | 51.0     | 60.3     | 78.5     | 780     |

### 4.2 Intensity correction methods analysis

In attempt to explain the reason why all intensity normalization steps was found redundant we hypothesise that by reducing data variance (reduce domain shifts) these methods also reduce signal variance. To test this idea we compute Kullback-Leibler divergence (KL) between healthy brain tissue and all tumor tissue. In almost all cases lower KL values corresponds to lower performance, e.g. for bias-field correction KL between healthy brain and WT tissue is equal to 0.47 compared to 0.61 of atlas registration, which coincide with segmentation quality drop (from 86.4 for atlas registration to 84.9 for bias-field corrected data). On the contrary, for denoized data KL are either the same or slightly larger compared to atlas data: 0.63 vs 0.61 for WT; 4.16 vs 4.01 for TC and 7.11 vs 6.70 for ET. Which completely coincides with segmentation performance. The only comparison which does not follow this explanations are bf-correction for TC (it
has lower KL compared to atlas data, but slightly better segmentation quality). Complete table with all KL comparisons are provided in the Appendix \[B\].

The overall results suggest the following recommendations:

– Image resampling is mandatory for accurate prediction, though even the fastest equalization of voxel size \([1, 1, 1]\) works fine.
– Bias-field correction, denoising and histogram matching seems to be redundant even for multi domain dataset. It can be explained by reduced data variance, and therefore reduced signal variance.
– Skull stripping helps model convergence, yet in terms clinically relevant measurements (difference in lesion volume measures), it has very small utility. Besides, it is the most time consuming step in the overall preprocessing pipeline.

5 Conclusion

We perform a rigorous ablation study of the most popular preprocessing steps used in the analysis of brain MRI images, including atlas registration, voxel resampling and image resizing, histogram matching, bias-field correction, denoising, and skull stripping.

We show that only image alignment and voxel resampling are essential for accurate prediction. We conclude that predictions after atlas registration do not significantly differ from ones with equal voxel resampling. We observe that bias-field correction, denoising, and histogram matching reduce data variance and do not affect DNN performance positively. We point out that skull stripping can lead to a measurable increase in accuracy and facilitate model convergence. On the other hand, brain extraction is very computationally expensive, and its incorporation into a pipeline does not affect clinically relevant volumetric measurements.

Thus we believe that skipping all steps excluding image alignment and voxel resampling from brain MRI deep learning pipeline may reduce computational costs and improve reproducibility across studies.

5.1 Work limitations

Our findings suggest that overall research reproducibility will benefit if one discards custom preprocessing steps, including different skull stripping, various implementations of bias field correction, denoising, etc.

However, the results can be contrary in transfer learning experiments, where data variance reduction can actually help model generalizability. This direction determines our future work on the topic.

Another extension could be using the more advanced DNN architectures for analysis, such as visual transformers.
References

1. Avants, B.B., Tustison, N., Song, G., et al.: Advanced normalization tools (ants). Insight j 2(365), 1–35 (2009)
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
3. Bakas, S., Zeng, K., Sotiras, A., Rathore, S., Akbari, H., Gaonkar, B., Rozycki, M., Pati, S., Davatzikos, C.: Glistrboost: combining multimodal mri segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In: BrainLes 2015. pp. 144–155. Springer (2015)
4. Beers, A., Gerstner, E., Rosen, B., et al.: Dicom-seg conversions for tcga-lgg and tcga-gbm segmentation datasets. Cancer Imaging Arch (2018)
5. Chang, K., Beers, A.L., Bai, H.X., Brown, J.M., Ly, K.I., Li, X., Senders, J.T., Kavouridis, V.K., Boaro, A., Su, C., et al.: Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. Neuro-oncology 21(11), 1412–1422 (2019)
6. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging 26(6), 1045–1057 (2013)
7. Eijgelaar, R.S., Visser, M., Müller, D.M., Barkhof, F., Vrenken, H., van Herk, M., Bello, L., Conti Nibali, M., Rossi, M., Sciortino, T., et al.: Robust deep learning-based segmentation of glioblastoma on routine clinical mri scans using sparsified training. Radiology: Artificial Intelligence 2(5), e190103 (2020)
8. Ermiş, E., Jungo, A., Poel, R., Blatti-Moreno, M., Meier, R., Knecht, U., Aebersold, D.M., Fix, M.K., Mauser, P., Reyes, M., et al.: Fully automated brain resection cavity delineation for radiation target volume definition in glioblastoma patients using deep learning. Radiation oncology 15(1), 1–10 (2020)
9. Győrfi, Á., Szilágyi, L., Kovács, L.: A fully automatic procedure for brain tumor segmentation from mult-spectral mri records using ensemble learning and atlas-based data enhancement. Applied Sciences 11(2), 564 (2021)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18(2), 203–211 (2021)
11. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., et al.: Automated brain extraction of multisequence mri using artificial neural networks. Human brain mapping 40(17), 4952–4964 (2019)
12. Kickingereader, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., Brugnara, G., Schell, M., Kessler, T., Foltyn, M., et al.: Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. The Lancet Oncology 20(5), 728–740 (2019)
13. Kondrateva, E., Pominova, M., Popova, E., Sharaev, M., Bernstein, A., Burnaev, E.: Domain shift in computer vision models for mri data analysis: an overview. In: Thirteenth International Conference on Machine Vision. vol. 11605, p. 116050H. International Society for Optics and Photonics (2021)
14. Menze, B., Isensee, F., Wiest, R., Wiestler, B., Maier-Hein, K., Reyes, M., Bakas, S.: Analyzing magnetic resonance imaging data from glioma patients using deep learning. Computerized medical imaging and graphics 88, 101828 (2021)
15. Moradmand, H., Aghamiri, S.M.R., Ghaderi, R.: Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. Journal of applied clinical medical physics 21(1), 179–190 (2020)
16. Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., et al.: Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430 (2018)
17. Nyúl, L.G., Udupa, J.K., Zhang, X.: New variants of a method of mri scale standardization. IEEE transactions on medical imaging 19(2), 143–150 (2000)
18. Pei, L., Bakas, S., Vossough, A., Reza, S.M., Davatzikos, C., Iftekharuddin, K.M.: Longitudinal brain tumor segmentation prediction in mri using feature and label fusion. Biomedical signal processing and control 55, 101648 (2020)
19. 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. Computer Methods and Programs in Biomedicine p. 106236 (2021). https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106236
20. de Raad, K., van Garderen, K., Smits, M., van der Voort, S., Incekar, F., Oei, E., Hirvasniemi, J., Klein, S., Starmans, M.: The effect of preprocessing on convolutional neural networks for medical image segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 655–658. IEEE (2021)
21. Ranjarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., Bendechache, M.: Brain tumor segmentation based on deep learning and an attention mechanism using mri multi-modalities brain images. Scientific Reports 11(1), 1–17 (2021)
22. Rathore, S., Bakas, S., Pati, S., Akbari, H., Kalarot, R., Sridharan, P., Rozycki, M., Bergman, M., Tunc, B., Verma, R., et al.: Brain cancer imaging phenomics toolkit (brain-captk): an interactive platform for quantitative analysis of glioblastoma. In: International MICCAI Brainlesion Workshop. pp. 133–145. Springer (2017)
23. Scarpace, L., Mikkelson, L., Cha, T., Rao, S., Tekchandani, S., Gutman, S., Pierce, D.: Radiology data from the cancer genome atlas glioblastoma multiforme [tcga-gbm] collection. The Cancer Imaging Archive 11(4), 1 (2016)
24. Smith, S.M., Brady, J.M.: Susan—a new approach to low level image processing. International journal of computer vision 23(1), 45–78 (1997)
25. Wang, G., Shapey, J., Li, W., Dorent, R., Dimitriadis, A., Bisdas, S., Paddick, I., Bradford, R., Zhang, S., Ourselin, S., et al.: Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 264–272. Springer (2019)
26. Wightman, R., Touvron, H., Jégou, H.: Resnet strikes back: An improved training procedure in timm. arXiv preprint arXiv:2110.00476 (2021)
A  Volumetric measurements

To estimate how different preprocessing steps affect clinically relevant measurements we compute MAE between true tumor volume (WT) and its’ estimate for each experiment. They are as follow:

- Resample to spacing - 11.8 ml
- Resample to size - 10.9 ml
- Atlas registration - 10.3 ml
- Atlas registration and bias-field correction - 12.7 ml
- Atlas registration and denoising - 10.4 ml
- Atlas registration and histogram matching - 10.3 ml
- Atlas registration and skull stripping - 9.2 ml
- Resample to spacing trained until convergence with data augmentations - 9.4 ml
- Atlas registration trained until convergence with data augmentations - 9.2 ml

With average tumor volume of 87 ml, the differences between the estimates are within 5% error margin for all experiments, and within 1% error margin for three best experiments. Additionally, we check if error in volume estimate depends on tumor volume Fig. 1. Surprisingly, we do not see any dependence.

![Fig. 1](image)

**Fig. 1.** Relation between tumor volume and its’ estimated volume from predicted mask for two experiments: Atlas registration and Atlas registration with skull stripping.

B  Skull stripping and registration

There are currently three main approaches to apply automatic skull stripping in multi-modal scenario. First, to register (via rigid registration) all modalities
to a single one (e.g. T1c), compute brain mask for this modality, and apply it on all others. Second, to compute brain mask for each MR sequence separately, apply the masks, and then register images to a single modality. Third, to register images without skull stripping and compute brain masks individually afterwards.

We performed experiment on GBM dataset to explore what type of HD-Bet skull-stripping to use in the main pipeline of experiments.

**Skull-stripping** We find out that the Dice score between skull-stripping, deployed before and after rigid registration is 0.958 (0.011), which means that rigid registration does not affect HD-Bet performance too much.

**Individual or shared mask for skull-stripping** Second, we compare individual brain mask with brain mask of T1c MR sequence (after rigid registration) to decide whether to use individual brain masks of shared one. Individual masks in most cases were smaller, than shared one (on average 13% in terms of volume). That implies potential loose of information on the borders on the skull, that can be critical to prediction. Thus, in all experiments we use shared mask (computed for T1c MR sequence) for all set of images. We use T1c MR sequence because in GBM dataset this modality has the largest resolution (on average).

## C Intensity equalization

We attempt to explain why popular intensity normalization steps have questionable effect on segmentation performance. Our hypothesis is that while they equalize modalities appearance across the data, they also reduce the differences between voxels’ intensities within each individual image. We compare intensity distribution for healthy brain voxels and voxels inside tumor mask using Kullback-Leibler divergence Tab. 2. We expect that if preprocessing step increases KL divergence, it should result in increased segmentation quality and vice versa. In most of the cases this supposition holds. Two exceptions are Denoising for Tumor core segmentation and Histogram matching for Enchancing tumor segmentation.

**Table 2.** KL divergence between healthy brain and tumor voxels. Lower values corresponds to smaller differences between healthy brain tissue and tumor tissue. All classes were compared to healthy brain (brain mask without Whole Tumor).

|                      | Whole Tumor | Tumor core | Enchancing tumor |
|----------------------|-------------|------------|------------------|
| Atlas registration   | 0.61(0.36)  | 4.01(3.62) | 6.70(3.53)       |
| Bias-field correction| 0.47(0.23)↑ | 3.92(3.70)↓ | 6.15(3.59)↑     |
| Denoising            | 0.63(0.37)↑ | 4.16(3.62)↑ | 7.11(3.59)↑     |
| Histogram matching   | 0.59(0.36)↑ | 3.94(3.63)↑ | 6.60(3.66)↓     |
Fig. 2. Visualization of preprocessing steps in ablation study.

D Preprocessing steps

We select several recent publications on brain MRI segmentation and identify most common preprocessing steps, see Table 3.

Table 3. Common preprocessing steps.

| Preprocessing step          | [9] | [21] | [18] | [14] | [8] | [7] | [22] | [25] | [2] | [12] |
|-----------------------------|-----|------|------|------|-----|-----|------|------|-----|------|
| Inter-modality registration | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Resample to spacing         | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Resample to size            | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Template registration       | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Histogram matching          | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Z-scoring                   | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Bias-field correction       | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Denoising                   | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |
| Skull stripping             | ✓   | ✓    | ✓    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   | ✓    |