REPRODUCIBILITY OF TUMOR SEGMENTATION OUTCOMES WITH A DEEP LEARNING MODEL

Morgane Des Ligneris, Axel Bonnet, Yohan Chatelain, Tristan Glatard, Michaël Sdika, Gaël Vila, Valentine Wargnier-Dauchelle, Sorina Camarasu-Pop, Carole Frindel

REPRODUCIBILITY OF TUMOR SEGMENTATION OUTCOMES WITH A DEEP LEARNING MODEL

Morgane des Ligneris [1], *Axel Bonnet* [1], *Yohan Chatelain* [2], *Tristan Glatard* [2], *Michaël Sdika* [1], *Gaël Vila* [1], *Valentine Wargnier-Dauchelle* [1], *Sorina Pop* [1,†] *and Carole Frindel* [1,†]

[1] Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, Lyon, France. [2] Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada.

ABSTRACT

In the last few years, there has been a growing awareness of reproducibility concerns in many areas of science. In this work, our goal is to evaluate the reproducibility of tumor segmentation outcomes produced with a deep segmentation model when MRI images are pre-processed (i) with two different versions of the same pre-processing pipeline, and (ii) by introducing numerical perturbations that mimic executions on different environments. Results show that these two variability sources can lead to important variations of segmentation outcomes: Dice can go as low as 0.59 and Hausdorff distance as high as 84.75. Moreover, both cases show a similar range of values, suggesting that the underlying causes for instability may be numerical stability. This work can be used as a benchmark to improve the numerical stability of the pipeline.

Index Terms— reproducibility, deep learning, numerical (in)stability, tumor segmentation

1. INTRODUCTION

Reproducibility has become a primary issue in research, particularly since data analysis workflows involve a large number of analysis steps that imply many possible choices. A recent study [1] underlines the huge variability in results when analyzing a single neuroimaging dataset by 70 different teams due to complex analysis workflows. Different configurations of a processing pipeline can induce up to 64% difference in the final segmentation produced by a trained deep learning model [2]. Such variations can seriously damage the confidence granted to the results. In clinical practice, it can be particularly problematic regarding consequences for patients because a change in results can lead to different (possibly wrong) diagnoses.

As shown by [3], everything matters when it comes to reproducibility, from the computational environment and software versions to the tool selection and the methodological approach. While tool selection and methodological approaches are tackled by recent studies such as [2] and [4], it is often neglected that the simple fact of using different operating systems may lead to non reproducible results as shown in [5]. This is particularly important for data pre-processing, which is one essential step prior to all kinds of data analysis. Public pre-processing pipelines are now used as a standard routine. Anyone can access and use those pipelines on their machines but little is known about the variability and consequences of the choice of the pipeline version or of the execution environment on the analysis, which is the focus of this paper.

In this paper, we evaluate the reproducibility of tumor segmentation outcomes of a deep learning model, DeepMedic [6, 7], after having pre-processed the data with the BRATS preprocessing pipeline [8, 9, 10] available in the Cancer imaging phenomics toolkit (CaPTk) [11, 12]. To this aim, we use the raw data from the publicly available dataset of Multiparametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania Health System (UPENN-GBM) [13, 14, 15]. The objective is to quantify the differences in results triggered by two factors: different versions of the same code and numerical perturbations that mimic executions on different environments, namely operating systems (OS).

2. MATERIAL AND METHODS

2.1. Dataset

The UPENN-GBM dataset is composed of 630 patients diagnosed with de novo GBM. For each patient, multi-parametric magnetic resonance imaging (mpMRI) scans are available including the four structural MRI scans: native T1-weighted (T1), post-contrast T1 (T1-GD), native T2-weighted (T2), and T2 fluid attenuated inversion recovery (T2-FL) scans. Among these 630 patients, we select[1] a subset of 191 complete patients and we use the four raw images provided in DICOM format as input.

† These authors have contributed equally to this work and share last authorship.

[1] Selection procedure in the GitLab : formating_dataset.ipynb

2.2. Pipeline

We use the BRATS-preprocess pipeline available in CaPTk. The pipeline has several steps[2] and intermediate files are saved. Studied files are represented in Figure 1 and referred to with different appellations, as described in the following. Raw data are designated by **'raw'**, and files after the reorientation to RAI (Right Anterior Inferior coordinate system), by **'rai'**. **'rai n4'** files correspond to the N4 biais correction, which is a temporary step helping with the registration to the atlas. It is not applied to the final images. The T1-GD image is registered to the atlas and then the other images (T1, T2 and T2-FL) to T1-GD. Images registered to the atlas space are named **'SRI'**. Following is the interpolation to a uniform isotropic resolution (1mm3) and the skull stripping, corresponding to **'brain'**. Finally, a **'segmentation'** is created through the DeepMedic algorithm. It contains three labels corresponding to different areas of the de novo Glioblastoma: necrosis, contrast-enhancing tumor, and edema.

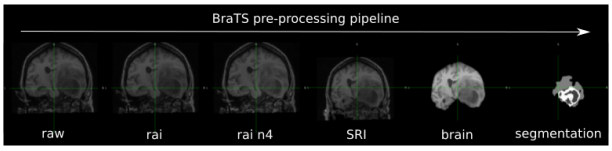


Fig. 1. Steps of BraTS pre-processing pipeline on the T1 image of patient UPENN-GBM-00002 in coronal view.

2.3. Experiments

In order to evaluate the difference in results triggered by the use of **different versions of the code**, we use versions v1.8.1 and v1.9.0 of the Brats pre-processing pipeline. More precisely, we use the Docker images[3] provided by the CBICA team with the image tags `CaPTk:2021.03.29` for v1.8.1 and `CaPTk:190rc` for v1.9.0. We first verify that results are deterministic and differences only come from the difference in pipeline versions. This requires that for multiple executions with the same inputs we obtain the same result when using the same Docker image (i.e., a given version of the pipeline). Two result files are considered the same if they have identical checksums[4]. We then compare the results obtained with the two versions of the pipeline on the same input files. More specifically, we compare the pre-processing steps and final mask of each result according to the metrics in section 2.5.

After the different versions of the pipeline, we evaluate the **influence of numerical perturbations**. As explained in [16] small amounts of noise or computational environments can lead to substantial differences in results. One important element of computational environments is operating systems (OS). "Fuzzy libmath" (FL) is a framework that uses Monte-Carlo arithmetic to simulate the variability induced by OS changes [17]. Here we evaluate the difference in results triggered by the change in OS as simulated with FL. Based on the `CaPTk` Docker images mentioned previously and on the FL Docker image and instructions[5], we built two new Docker images corresponding to the studied versions of the pipeline (v1.8.1 and v1.9.0) instrumented with FL. For each of these versions, we executed three repetitions for each subject.

2.4. Execution environment

All BraTS experiments presented in the paper are executed on the Virtual Imaging Platform (VIP) [18]. VIP is a web portal for medical simulation and image data analysis. By effectively leveraging the computing and storage resources of the EGI federation[6], VIP offers its users high-level services enabling them to easily execute medical imaging applications on a large scale computing infrastructure. While in production conditions VIP uses distributed and heterogeneous resources, all experiments presented in this paper were run under controlled conditions. The experiments comparing the two pipeline versions were executed on a single VM hosted at CREATIS. Due to the multiple repetitions (and hence longer computing time) required by the experiments using FL, they were all executed on three identical VMs hosted on the SCIGNE platform[7].

2.5. Evaluation

Ideally, results obtained with the same inputs should be identical. Since this is unfortunately not always the case, differences in results are quantified using different well-known metrics as described below. We use the md5 checksum function to assess if files are identical between two executions. This verification is done for all the files, from 'raw', to the final step 'brain'. To evaluate the differences in the 'segmentation', we use the Sørensen–Dice Coefficient (Dice) as an overlap-based metric and the Hausdorff Distance (HD) as a boundary-based metric. To explore differences between files earlier in the process, before the registration to the same space, we need metrics able to quantify the similarity between two images, such as Peak Signal to Noise Ratio (PSNR). For the experiment using FL, to evaluate changes in precision for intermediate files we use the significant digit metrics. Similarly to what is done in [17], we measure results precision as the number of significant bits among result samples obtained with the "fuzzified" pipeline, as $s = -\log_2 \left| \frac{\sigma}{\mu} \right|$, where σ and μ are respectively the observed cross-sample standard deviation and average.

[2] `https://cbica.github.io/CaPTk/preprocessing_brats.html`

[3] `https://hub.docker.com/r/cbica/captk`

[4] `https://en.wikipedia.org/wiki/Md5sum`

[5] `https://github.com/verificarlo/fuzzy`

[6] `https://www.egi.eu/egi-federation/`

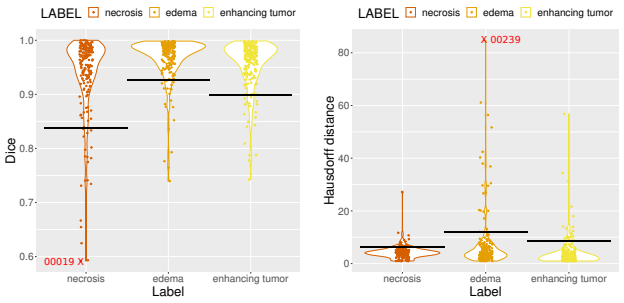
[7] `https://scigne.fr/en/page-dacceuil-english/`

3. RESULTS

3.1. Impact of different versions of the pipeline

Checksums are identical for two executions of the same pipeline version on VIP (see on the GitLab checksums.ipynb[8]). This confirms that results are deterministic and that the differences we may observe in the following only come from the difference in pipeline versions.

To evaluate the impact of the two different versions of the pipeline, Figures 2-a and -b respectively depict the Dice and Hausdorff metrics computed for the pairs of results of the two versions on all the 191 subjects. Scores are computed for each label: necrosis, edema, and contrast-enhancing tumor. The mean and standard deviation (SD) for the two metrics are given in table 1. We notice that even if mean values for Dice are above 0.94 (and HD below 7.37), outliers can go as down as 0.59 (and up to 84.75 for HD). As shown with the black line, respectively for the necrosis, edema, and contrast-enhancing tumor label, 10% of the patients have Dice under 0.84, 0.93, and 0.90 (and HD larger than 6.4, 12.08, and 8.54). We observe more Dice outliers for the label necrosis (Figure 2-a) and more HD outliers for the label edema (Figure2-b). These differences can be explained by the fact that necroses are more often small structures inside the contrast-enhancing tumors, while edema can be stretched all around. Figures 2-c and -d show tumor segmentations and the brain images for the two patients highlighted in red, allowing to better understand where the differences lie. They are also representative of the other outliers. The necrosis segmentation for patient 00019

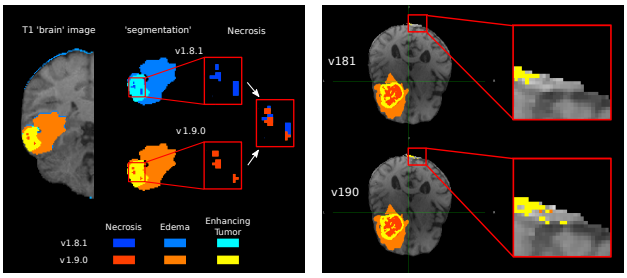


(a) Dice (b) Hausdorff Distance

Fig. 2. Evaluation of segmentation results with (a) Dice and (b) HD for each label between v1.8.1 and v1.9.0.

(Fig.2-a) has a Dice of **0.59** and HD of 6.08. This low Dice is the consequence of different registration results on the SRI atlas between v1.8.1 (in blue) overlaid with v1.9.0 (in yellow, orange or red). Since there are few pixels labelled as necrosis (in red), the misregistration has a higher impact on the Dice value for this label. The edema segmentation for patient 00239 (Figure 2-b) has a Dice of 0.98 and HD of **84.75**. The

[8] https://gitlab.in2p3.fr/MDL/reprovip-wp3-use-case/-/blob/master/metrics/checksums.ipynb



(a) UPENN-GBM-00019 (b) UPENN-GBM-00239

Fig. 3. Patients in coronal view with T1 'brain' image in grey tones (a) On the half brain image, the results of v1.9.0 are superimposed on those of v1.8.1. On the right, the segmentation and necrosis details for both versions are shown first, and finally the overlay between the versions is only for necrosis. (b) v1.8.1 at the top and v1.9.0 at the bottom with 'segmentation' overlaid on top of T1 image with label necrosis in red, contrast-enhancing tumor in yellow, and edema in orange. Zoom to visualize pixels of edema detected in v1.9.0 but not in v1.8.1.

high value of HD associated with edema can be explained in the zoomed area where pixels are associated with edema (in orange) for v1.9.0 but not for v1.8.1. With v1.9.0 the edema label is wrongly associated with a few pixels located far away from the core of the edema around the contrast-enhancing tumor, thus explaining the high HD score.

For a deeper understanding of reproducibility issues, we also examine the intermediate files produced by the pipeline. Indeed, as we detect differences between versions for the segmentation task, we study, through the intermediate files, where variability occurs and how it propagates through the pipeline. PNSR values are depicted in Figure 4. The larger

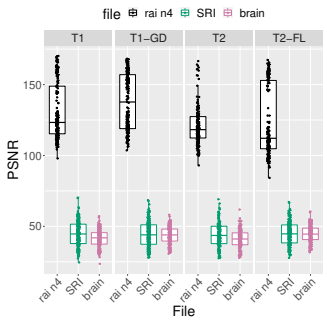


Fig. 4. PSNR for intermediary files from step rai n4 to the skull stripping between v1.8.1 and v1.9.0.

the value, the greater the similarity between the two files. The PSNR for the first two steps ('raw' and 'rai') is not represented, because it is equal to infinity, attesting that the images were identical between the versions. For the 'rai n4' step, the PSNR values are high and scattered, showing that the N4 bias correction is the first step introducing variations. For the following steps, the PSNR value decreases sharply and continues until the last step 'brain'.

	v1.8.1 VS v1.9.0			Fuzzy-Libmath		
Dice	Mean	SD	Min	Mean	SD	Min
Nec	0.94	0.08	0.59	0.93	0.09	0.0
Ed	0.97	0.04	0.74	0.96	0.04	0.66
CET	0.96	0.05	0.74	0.95	0.05	0.54
HD	Mean	SD	Max	Mean	SD	Max
Nec	4.31	2.47	27.20	4.31	3.28	40.62
Ed	7.36	11.03	84.75	7.71	10.52	84.75
CET	4.26	5.79	56.82	4.37	7.09	69.55

Table 1. Mean and SD for Dice and HD between v1.8.1 and v1.9.0 and for FL runs. Nec: Necrosis, Ed: Edema and CET: Contrast-Enhancing Tumor

3.2. Impact of numerical perturbations

We now evaluate the differences in results due to the numerical perturbations introduced with FL. For each version instrumented with fuzzy (v1.8.1 fuzzy and v1.9.0 fuzzy), we executed 3 repetitions and then computed all Dice coefficients between each FL execution and the corresponding pipeline version without FL. In Table 1, we can see the mean and SD for the two experiments: the change of version and the simulated change of OS with FL. We notice that the order of magnitude is very similar.

We further investigate which step of the pre-processing pipeline introduced variations by computing the the mean significant digits. Results of significant digits[9] are only illustrated for v1.8.1 with FL in Figure 5 since results for v1.9.0 are very similar. Similarly to PSNR outcomes, the first two

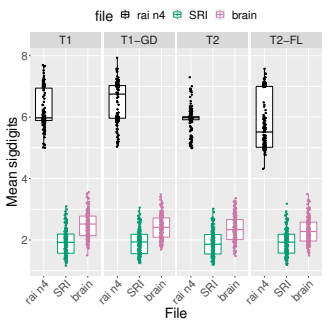


Fig. 5. Mean significant digits v1.8.1 with Fuzzy-Libmath

steps ('raw' and 'rai') produce identical images (mean significant digit of exactly 8) and are not represented. We can observe that there is a drop in similarity at the N4 bias correction step. The loss of similarity is accentuated in the next step 'SRI', and is also very small for 'brain'. The slight increase in significant digits for the last step 'brain' may come from the skull stripping involving the deletion of pixels and replacing them with background (more significant digits of 8) and so reducing the number of pixels with low significant digits and increasing their mean value.

[9] https://github.com/glatard/fuzzy-fmriprep/blob/main/sigdigits.ipynb

4. DISCUSSION AND CONCLUSION

In this paper, we explored the degree of reproducibility of tumor segmentation results via the inference of a deep learning model when using different versions of a pre-processing pipeline and digital perturbations that mimic executions in different environments. Our experiments highlight the fact that the results are not reproducible bit by bit. We quantified the differences using several complementary metrics, such as the Sørensen-Dice coefficient, Hausdorff distance for the final segmentation results and peak signal-to-noise ratio, the number of significant digits for the intermediary images in the pre-processing pipeline.

Based on the results presented in section 3, the main take-home messages are:

1. There is an important variation of segmentation outcomes between versions of the BRATS pipeline. Even though on average Dice coefficients are high, values can go down to 0.59 (which is very low) and 10% of patients for the label necrosis are under 0.84.
2. The inter-OS variability measured with FL is in the same order of magnitude as the between-version variability, which suggests that the underlying causes for instability may be numerical stability.
3. N4 normalization and SRI seem to be the main steps in the pipeline where most of the variability comes from.
4. The variability in segmentation outcomes depends on the input data. This data sensitivity may introduce a bias in model performance at the patient level.

In conclusion, we believe it is important to review the numerical stability of the pipeline. Reproducibility experiments like the one presented in this paper can be used as a benchmark to improve it.

These are the first results of a larger study within the ReproVIP project. Future work includes: (i) further analysis of the source of variability among the two versions of the pipeline, (ii) experiments allowing to quantify and compare the variability introduced by different computing infrastructures (we are planning to use the Grid5000[10] experimental platform), (iii) re-train the model and evaluate the impact on the model transfer.

5. DATA AVAILABILITY STATEMENT

The project code is available in the following GitLab project reprovip-wp3-use-case.

6. CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

[10] https://www.grid5000.fr/w/Grid5000:Home

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access from the publicly available repository of The Cancer Imaging Archive at `https://doi.org/10.7937/TCIA.709X-DN49`. Ethical approval was not required as confirmed by the license attached with the open access data.

8. FUNDING

This work was supported by the French ANR through the ReproVIP project (ANR-21-CE45-0024-01).

9. ACKNOWLEDGMENTS

The authors acknowledge the support of F. Bellet, J. Pansanel and A. Tsaregorodtsev for code deployment on resources provided by CREATIS and the SCIGNE platform using the VIP portal and the Dirac service, as well as the support and resources provided by France Grilles and the biomed VO of the EGI infrastructure. This work was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063). The authors would also like to thank CaPTk developers for their help and fruitful exchanges.

10. REFERENCES

[1] R. Botvinik-Nezer, F. Holzmeister, C. Camerer, et al., "Variability in the analysis of a single neuroimaging dataset by many teams," 11 2019.

[2] K. De Raad, K. A. van Garderen, M. Smits, et al., "The effect of preprocessing on convolutional neural networks for medical image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 655–658.

[3] D. N. Kennedy, S. A. Abraham, J. F. Bates, et al., "Everything matters: The repronim perspective on reproducible neuroimaging," *Frontiers in Neuroinformatics*, vol. 13, pp. 1, 2019.

[4] A. Bowring, C. Maumet, and T. E. Nichols, "Exploring the impact of analysis software on task fmri results," *Human brain mapping*, vol. 40, no. 11, pp. 3362–3384, 2019.

[5] T. Glatard, L. B. Lewis, R. Ferreira da Silva, et al., "Reproducibility of neuroimaging analyses across operating systems," *Frontiers in Neuroinformatics*, vol. 9, pp. 12, 2015.

[6] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, et al., "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[7] K. Kamnitsas, L. Chen, C. Ledig, et al., "Multiscale 3d convolutional neural networks for lesion segmentation in brain mri," *Proc. MICCAI Ischemic Stroke Lesion Segmentation Challenge*, 01 2015.

[8] B. H. Menze, A. Jakab, S. Bauer, et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[9] S. Bakas, H. Akbari, A. Sotiras, et al., "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, 09 2017.

[10] S. Bakas, M. Reyes, A. Jakab, et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," 11 2018.

[11] C. Davatzikos, S. Rathore, S. Bakas, et al., "Cancer imaging phenomics toolkit: Quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome," *Journal of Medical Imaging*, vol. 5, no. 1, Jan. 2018.

[12] S. Pati, A. Singh, S. Rathore, et al., "The cancer imaging phenomics toolkit (captk): Technical overview," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds., Cham, 2020, pp. 380–394, Springer International Publishing.

[13] M. Elkjær, M. Andersen, S. Hoyer, et al., "Multiparametric magnetic resonance imaging monitoring patients in active surveillance for prostate cancer: a prospective cohort study," *Scandinavian Journal of Urology*, vol. 52, pp. 1–6, 12 2017.

[14] S. Bakas, C. Sako, H. Akbari, et al., "The university of pennsylvania glioblastoma (upenn-gbm) cohort: advanced mri, clinical, genomics, amp; radiomics," *Scientific data*, vol. 9, no. 1, pp. 453, July 2022.

[15] K. Clark, B. Vendt, K. Smith, et al., "The cancer imaging archive (tcia): Maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, Copyright: Copyright 2013 Elsevier B.V., All rights reserved.

[16] G. Kiar, Y. Chatelain, P. de Oliveira Castro, et al., "Numerical uncertainty in analytical pipelines lead to impactful variability in brain networks," *PloS one*, vol. 16, no. 11, pp. e0250755, 2021.

[17] A. Salari, Y. Chatelain, G. Kiar, and T. Glatard, "Accurate simulation of operating system updates in neuroimaging using monte-carlo arithmetic," 08 2021.

[18] T. Glatard, C. Lartizien, B. Gibaud, et al., "A virtual imaging platform for multi-modality medical image simulation," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 110–118, 2013.