Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059000
Article Type:	Original research
Date Submitted by the Author:	05-Nov-2021

Complete List of Authors:	Gryska, Emilia; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Medical Radiation Sciences, Institute of Clinical Sciences Björkman-Burtscher, Isabella; University of Gothenburg Sahlgrenska Academy, Department of Radiology, Institute of Clinical Sciences; Sahlgrenska University Hospital, Department of Radiology Jakola, Asgeir Store; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology; Sahlgrenska University Hospital, Department of Neurosurgery Dunås, Tora; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology; University of Gothenburg Sahlgrenska Academy, Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology Schneiderman, Justin; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology Heckemann, Rolf; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Medical Radiation Sciences, Institute of Clinical Sciences

Keywords:	Magnetic resonance imaging < RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING, Neuroradiology < RADIOLOGY & IMAGING, STATISTICS & RESEARCH METHODS

Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study

Emilia A. Gryska[1,2*], Isabella M. Björkman-Burtscher[3,4], Asgeir S. Jakola[5,6], Tora Dunås[5,7], Justin F. Schneiderman[1,5], Rolf A. Heckemann[1,2]

1 MedTech West at Sahlgrenska University Hospital, University of Gothenburg, Gothenburg, Sweden

2 Department of Medical Radiation Sciences, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

3 Department of Radiology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

4 Department of Radiology, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden

5 Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

6 Department of Neurosurgery, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden.

7 Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Sweden

* Corresponding author: Emilia A. Gryska

Postal address: MedTech West, Röda stråket 10B, Sahlgrenska University Hospital, 413 45 Gothenburg, Sweden

E-mail address: emilia.gryska@gu.se

Telephone number: +46 720304106

Keywords: Reproducibility of Results, Deep Learning, Brain Neoplasms, Magnetic Resonance Imaging

Word count: 3 980

ABSTRACT

Objectives: To determine the reproducibility and replicability of studies that develop and validate segmentation methods for brain tumours on MRI and that follow established reproducibility criteria; and to evaluate whether the reporting guidelines are sufficient.

Methods: Two eligible validation studies of distinct DL methods were identified. We implemented the methods using published information and retraced the reported validation steps. We evaluated to what extent the description of the methods enabled reproduction of the results. We further attempted to replicate reported findings on a clinical set of images acquired at our institute consisting of high and low grade glioma (HGG, LGG), and meningioma (MNG) cases.

Results: We successfully reproduced one of the two tumour segmentation methods. Insufficient description of the preprocessing pipeline and our inability to replicate the pipeline resulted in failure to reproduce the second method. The replication of the first method showed promising results in terms of Dice similarity coefficient (DSC) and sensitivity (Sen) on HGG cases (DSC=0.77, Sen=0.88) and LGG cases (DSC=0.73, Sen=0.83), however poorer performance was observed for MNG cases (DSC=0.61, Sen=0.66). Preprocessing errors were identified that contributed to low quantitative scores in some cases.

Conclusions: Established reproducibility criteria do not sufficiently emphasize description of the preprocessing pipeline. Discrepancies in preprocessing as a result of insufficient reporting are likely to influence segmentation outcomes and hinder clinical utilization. A detailed description of the whole processing chain, including preprocessing, is thus necessary to obtain stronger evidence of the generalizability of DL-based brain tumour segmentation methods and to facilitate translation of the methods into clinical practice.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This is an independent evaluation of the reproducibility of DL-based lesion segmentation studies that follow established reporting guidelines.
- We assessed existing reproducibility checklists and developed an update proposal emphasizing the need for detailed reporting of preprocessing.
- The clinical data set acquired at our institution was suitable for the replication part of the study.
- This study did not aim to enable inferences about the clinical utility of the evaluated algorithms.

INTRODUCTION

The scientific community has directed substantial efforts at developing deep-learning (DL) methods for medical image analysis. DL methods have become the default choice in automatic medical image analysis under the claim of superior performance to classical algorithms.[1-3] However, their outstanding performance comes at the cost of high complexity and inherent variability in model performance.[3] Consequently, assessing which model design choices determine the empirical gains is challenging.[3-5] Critics have also pointed out that scientific reporting of study designs has often been insufficient, and that the analysis of results tends to be biased towards authors' desired outcomes.[4, 6, 7] These issues present critical challenges to realizing the potential of AI and translating promising scientific findings into reliable and trusted evidence-based medicine.

The problem has been recognized by researchers, and efforts have been made to standardize reporting practices of DL validation studies. The checklist proposed by Pineau et al.[6, 8] identifies a set of items to be reported pertaining to the presented models/algorithms, theoretical claims, data sets, code, and experimental results. The negative impact of publishing non-reproducible studies has prompted the Medical Image Computing and Computer Assisted Intervention (MICCAI) Society[9] to adopt the checklist. Authors submitting manuscripts to MICCAI conferences are now required to complete it upon submission.

The reproducibility problem in relation to the specific field of medical image segmentation was highlighted by Renard et al. in a literature review.[3] The authors reviewed studies describing DL segmentation methods to determine whether there is "enough information in published articles […] to correctly reproduce the results".[3] The authors only found three out of twenty-nine studies to be sufficient in this regard. While no specific reproducibility checklist has been defined particularly for medical image segmentation, the authors present recommendations for the framework description that provides specific context for this task. Their recommended items to be reported[3] are largely congruent with those proposed by Pineau et al.[6, 8] Renard et al.,[3] however, group their items by sources of variability in the model and evaluation framework, in contrast to grouping by scientific article section, as originally proposed by Pineau et al.[6, 8]

Two of the three methods that Renard et al. identify as reproducible are DL-based models for brain tumour segmentation.[10, 11] This particular task receives a lot of attention from the medical image analysis community.[12] In our scoping review of automatic methods for brain lesion segmentation, we found issues with reporting that may affect reproducibility.[5] In particular, reporting of the preprocessing steps is inadequate in many instances. Preprocessing, while mentioned in both Pineau's checklist[6, 8] and Renard et al.'s[3] recommendations, does not receive the emphasis and detail that is given to other parts of the framework description.

The aim of this study was to determine the reproducibility and replicability of the two methods for brain tumour segmentation[10, 11] that Renard et al. identified as adequately reported;[3] and to evaluate whether Renard's and Pineau's reproducibility recommendations are sufficient also for the task to segment an in-house clinical data set of brain tumours.

MATERIAL AND METHODS

Overview

The study design is based on the assumption that the reproducibility items proposed by Renard et al. are necessary and sufficient for reproduction and replication. We attempted to reproduce (according to the definition from the National Academies of Sciences, Engineering and Medicine[13], also used by Pineau et al.[6]) the two methods for brain lesion segmentation[10, 11] that Renard et al. identified as adequately reported.[3] Our goal was to implement the respective original methods with all processing steps and parameters and test them on the same data on which they were originally validated (reproducibility). As a measure of success, we compared quantitative results on segmentation accuracy to those reported in the original studies. We then attempted to replicate[6, 13] the findings: we performed an external validation on a clinically obtained data set from our institution.

Patient and Public Involvement

No patient involved.

Reproducibility analysis

Evaluated segmentation algorithms

We implemented the two previously proposed DL algorithms for brain tumour segmentation: DeepMedic by Kamnitsas et al.[10] and an algorithm proposed by Pereira et al.,[11]. In Table 1 these algorithms are described in compliance with the reproducibility categories listed by Renard et al.,[3] together with libraries and computational parameters we used in our implementations. We trained DeepMedic and tested both algorithms on a cluster with a Tesla V100 GPU (5120 cores; Nvidia Corp., Santa Clara, CA, USA), 32 GB RAM, and two 8-core Xeon Gold 6244 @ 3.60GHz processors (Intel Corp., Santa Clara, CA, USA).

In the following subsections *DeepMedic* and *Pereira et al.,* we describe the input data requirements and preprocessing procedures that we deemed necessary for reproduction of the algorithms that go beyond Renard et al.'s recommendations.

DeepMedic

The authors of the DeepMedic study[10] made the software available for independent evaluation (https://github.com/deepmedic) but did not provide a trained model. The software came with a set of configurable network parameters and requirements for the input data. The input data requirements were: images in NIfTI[21] format; images for each patient and reference labels with optional brain tissue masks (regions of interest – ROIs) had to be co-registered; all images fed to the network had to have the same voxel size; and for optimal performance, MR signal intensities had to be standardized to have zero-mean and unit-variance within each ROI.

Pereira et al.

Pereira et al. published the two network architectures (HGG – high grade glioma and LGG – low grade glioma) with trained weights,[14] both of which we used to reproduce the validation. No other processing was included in the published code; independent setup of the processing pipeline was therefore required. The preprocessing described in the original publication consisted of bias field correction with N4ITK,[22] followed by intensity normalization[23] of each imageThe input patch intensities were finally normalized with the mean and standard deviation calculated from the training patches across each sequence. A roughly similar number of patches was extracted for each class (approximately 50 000 per class for HGG to match the number of patches extracted for training as stated in the original article). The segmentation result was further processed by removing clusters of voxels smaller than a predefined threshold of 10 000 mm^3 and 3 000 mm^3 in HGG and LGG, respectively.

Table 1: Description of the two algorithms implemented in the reproducibility analysis, DeepMedic[10] and Pereira et al.'s,[11] according to the reproducibility categories proposed by Renard et al.[3] and Python version and libraries used for our implementations. CNN – convolutional neural networks, CRF – conditional random field, CV – cross-validation, DSC – Dice similarity coefficient, FC – fully connected, HGG – high grade glioma, LGG – low grade glioma.

Main category	Sub-category	DeepMedic	Pereira et al.
Algorithm/model	Description of the DL architecture	Dual-path 3D CNN with a fully connected 3D CRF.[15]	Single-path 2D CNN; two network architectures for HGG and LGG.
Dataset description	Image acquisition parameters Image size Data set size Link to the data set	BraTS 2015 dataset[16]	
Preprocessing description	Data excluded + reason	none	none
	Augmentation transformation	Sagittal reflection of images	Rotation with multiples of 90° angles
	Final sample size	Not specified	~1 800 000 for HGG ~1 340 000 for LGG
Training/validation/ testing split	Explanation if validation set not created	Training and testing sets provided by the BraTS challenge	
CV strategy + number of folds	Not specified	5-fold CV on training set	1 subject in both HGG and LGG
Optimization strategy	Optimization algorithm + reference	RMSProp optimizer[17] and Nesterov's momentum[18]	Stochastic Gradient Descent and Nesterov's momentum[18]
	Hyperparameters (learning rate *a*, batch size *n*, drop-out *d*)	a = 10-3 (halved when the convergence plateaus); n = 10 d = 50% (int he last 2 hidden layers)	a_{inital} = 0.003 a_{final} = 0.00003 n = 128 $dHGG$ = 0.1 (in FC layers) $dHGG$ = 0.5 (in FC layers)
	Hyperparameter selection strategy	CRF: 5-fold CV on a training subset	Validation using 1 subject in both HGG and LGG
Computing infrastructure	Name, class of the architecture, and memory size	NVIDIA GTX Titan X GPU using cuDNN v5.0, 12GB	GPU NVIDIA GeForce GTX 980
Middle-ware	Toolbox used/in-house code + build version	Theano[19]	Theano [19], Lasagne[20]
	Source code link + dependencies	https://github.com/deepmedic	http://dei-s2.dei.uminho.pt/pessoas/csilva/brats_cnn/
Evaluation	Metrics average + variations	Mean of DSC, Precision, and Sensitivity (calculated by the online evaluation system)	Boxplot and mean of DSC (calculated by the online evaluation system)

Our implementation middleware		
Python version	3.8.2	3.7.4
DL library	Tensorflow 2.2.1	Theano (git version eb6a412), Lasagne (git version 5d3c63c)
Numpy	1.18.5	1.17.3
Nibabel	3.0.2	3.2.1

Image data set used for reproducibility analysis

Both algorithms were originally validated in the 2015 Brain Tumor Segmentation Challenge (BraTS),[24] which consists of training and testing image sets of patients diagnosed with HGG and LGG. The training set contains 274 examinations (HGG n = 220, LGG n = 54). Each examination consists of T1-weighted (T1w) images before and after injection of contrast material (CM), T2w, and FLAIR (fluid-attenuated inversion recovery) images. The training data set additionally contains manual segmentations of tumour structures that serve as a criterion standard and delineate necrotic core, contrast-enhancing (CE) core, non-CE core, and oedema. For the test set containing 110 examinations the criterion standard segmentations are not publicly available.Users can upload their segmentation results to an online system[16, 25] that internally compares the results with the hidden reference to determine per-case metrics (Dice similarity coefficient – DSC, positive predictive value – PPV, sensitivity, and kappa). The system then returns summary measures (means and ranking position) to the user. Images in both sets are provided in .mha format and have been preprocessed with spatial normalization,[26] skull-stripping,[27] and resampling to an isotropic resolution of 1 mm^3 (linear interpolator).

Outcome parameters

We experimentally evaluated whether the two methods that Renard et al. identified as reproducible according to their proposed criteria[3] were possible to reproduce. Specifically, we examined whether enough information was given in the original articles or supplementary information for each processing step. If re-implementation did not reproduce the originally reported results, we contacted the authors directly to follow up on any missing details and added this information to the results. Pereira et al. supplied a pre-trained model;[14] for DeepMedic, we trained our re-implementation on the BraTS 2015 training data. Thereafter, we segmented the BraTS 2015 test set with both methods. We submitted the resulting segmentations to the online evaluation system[16] and recorded the summary measures returned (mean DSC, mean sensitivity, and mean PPV). Finally, we compared the summary measures with those available in the original publications.

Replication analysis

Evaluated segmentation algorithm

Only DeepMedic was successfully re-implemented (cf. Results – Reproducibility study). External validation (replication analysis) on in-house clinical data was therefore carried out with DeepMedic. The segmentation models trained on the BraTS training data in the reproducibility analysis were applied to our dataset using a workstation with an Intel Core i7-6700HQ CPU @ 2.60 GHz processor and Nvidia GTX960M graphics card.

Image data set used for the replication analysis

The clinical in-house testing data set consisted of images from 27 cases (HGG n = 12; LGG n = 10; meningioma – MNG n = 5). The set was selected for this study from a larger sample of image data. Data were pseudonymized and inclusion criteria were pre-operative examinations, availability of manual expert reference segmentations, and imaging findings typical for the included types of pathology.

As in the BraTS data set, each MR examination included non-CM T1w, CM T1w, T2w, and FLAIR images. The images were provided in NIfTI[21] format. Since we used a model trained on BraTS data to segment these images, we used the BraTS-Processor module from the BraTS Toolkit[28] for preprocessing. Binary lesion segmentations had been prepared by trained personnel and revised by a senior neurosurgeon (AJ). Whole-tumour labels generated by delineation of T2/FLAIR hyperintensities were used for LGG. For HGG and MNG, the tumour core label was used, which had been delineated on CM T1w images and included CE tumour as well as any components enclosed by CE tumour. The reference segmentations were registered from the native space to the BraTS space following the transformation steps and using the registration matrices generated by the BraTS-Processor.[28]

Outcome parameters

The replicability of DeepMedic was assessed by comparing DSC, sensitivity, and PPV derived from processing the clinical in-house data with those provided by the online system[16] during the reproducibility analysis on the BraTS test set. DSC, sensitivity, and PPV were evaluated for the whole tumour label generated by the algorithms for LGG cases and the tumour core label for HGG and MNG cases. We visually evaluated individual cases to determine causes of segmentation errors.

Based on findings from the reproducibility and the replication analysis we reviewed recommendations on reporting items proposed by Renard et al.,[3] and Pineau et al.[8] Challenges and failures in our attempts at reproduction and replication were documented and examined throughout the processes above. We then assessed and summarized these outcomes with suggested specific improvements to the reproducibility items for lesion segmentation on magnetic resonance images for brain segmentation.

RESULTS

Reproducibility study

DeepMedic

BraTS data fulfilled most of the input requirements for DeepMedic, apart from the format and the image intensity normalization. To reproduce the study, all images were converted to NIfTI format, and MR signal intensities were normalized to have zero-mean and unit-variance within each ROI. We implemented these steps using SimpleITK for image conversion and an in-house python program for signal intensity normalization. Since the BraTS images are already skull-stripped, we generated brain masks for each patient by thresholding each image to include only non-zero voxels in order to reduce the runtime of the algorithm. The only changes we made in the DeepMedic configuration file were to set the number of input channels to all four available, as described in the original article (default in the source code was CE T1w and FLAIR), and to specify not to perform validation of the available samples, as the hyperparameters had already been defined for the model. Training DeepMedic took approximately 27 hours, and testing took 14.5 minutes.

The quantitative evaluation shows that our re-implementation and testing of DeepMedic on the BraTS 2015 data set achieved comparable results to those presented in the original study (Table 2). We therefore deem the method reproducible.

Table 2: Reproducibility results on BraTS 2015 presented in the original paper for DeepMedic[10] and for Pereira et al.'s method[11] (original) and for our independent reproducibility analysis (this work). Our analysis was carried out for high grade glioma (HGG) and low grade glioma (LGG) model parameters of the Pereira et al.'s method. The results were congruent with the original analysis for DeepMedic but they show an unsuccessful attempt to reproduce Pereira et al.'s work. The higher score in each column is emphasized in bold. Measures of dispersion or significance of differences were not available for the original method evaluation. CE – contrast-enhanced.

	Dice similarity coefficient			**Positive predictive value**			**Sensitivity**		
	Whole tumour	**Tumour core**	**CE tumour**	**Whole tumour**	**Tumour core**	**CE tumour**	**Whole tumour**	**Tumour core**	**CE tumour**
DeepMedic									
Original	0.85	0.67	0.63	0.85	**0.85**	**0.63**	0.88	0.61	0.66
This work	0.85	**0.68**	**0.64**	0.85	0.83	0.62	0.88	**0.64**	**0.70**
Pereira et al.									

Original	**0.78**	**0.65**	**0.75**	-	-	-	-	-	-
This work (HGG)	0.36	0.25	0.17	0.36	0.21	0.29	0.54	0.58	0.17
This work (LGG)	0.25	0.14	0.13	0.40	0.51	0.37	0.25	0.10	0.10

Pereira et al.

The preprocessing description by Pereira et al. lacked certain parameters pertaining to the intensity normalization: percentile points used to create a reference histogram for each sequence and glioma grade, and intensity parameters of the training patches. Furthermore, it was not specified which model architecture was used on the BraTS 2015 test set, where the data include both HGG and LGG. Despite the missing parameters, we made an attempt to reproduce the study. We used N4ITK bias field correction (as implemented in SimpleITK) with default parameters and a histogram normalization procedure adapted from Reinhold et al.[29] We decided on this implementation instead of the corresponding function in SimpleITK, because the latter requires a reference image or histogram, neither of which was available. For the final patch-normalization step, the intensity parameters were not available, so we normalized each test image ROI to have zero-mean and unit-variance. Finally, the results were post-processed according to the procedure described by the authors. The testing time of Pereira et al.'s method was approximately 8 hours.

As the attempt was unsuccessful (results of the quantitative evaluation presented in Table 2), we approached the lead author of the method and requested the missing information. The author generously provided information on the bias field correction as well as image histogram normalization parameters.

Following this input, the N4ITK bias field correction was conducted using the implementation in ANTs[30] with the wrapper in Nipype[31] with the following parameters specified: *n_iterations* = [20, 20, 20, 10], *dimension* = 3, *bspline_fitting_distance* = 200, *shrink_factor* = 2, *convergence_threshold* = 0. A visual inspection of the field inhomogeneity correction with ANTs/Nipype and the parameters given versus SimpleITK showed signal intensity differences in the tumour region (Figure 1) that plausibly explained the failure to reproduce.

The implementation of Nyul's algorithm[23] for intensity normalization was developed in the lead author's former lab, and the author was not at liberty to share the code. Instead, the author provided percentile points and corresponding intensity landmarksfor each MR sequence used in their implementation. In the original study, however, the authors trained separate sets of parameters for LGG and HGG and could not retrieve the patch intensity parameters for patch normalization. At this point, we decided not to pursue further efforts to reproduce the study.

Replication analysis

The replication analysis was conducted on DeepMedic only. Quantitative results of the comparison of automatic segmented MR images collected in-house and expert delineations of the chosen tumour labels are presented in Table 3.

Table 3: DeepMedic[10] replication analysis results on in-house data for high grade glioma (HGG) cases and meningioma (MNG) cases evaluated on the tumour core and for low grade glioma (LGG) cases evaluated on the whole tumour label. DSC – Dice similarity coefficient, PPV – positive predictive value, Sen – sensitivity, Std. – standard deviation.

ID	01	02	03	04	05	06	07	08	09	10	11	12	Mean	Std.
HGG cases tumour core														
DSC	0.88	0.85	0.80	0.85	0.89	0.85	0.57	0.89	0.86	0.81	0.87	0.14	**0.77**	**0.22**
PPV	0.84	0.86	0.72	0.84	0.85	0.79	0.41	0.85	0.80	0.73	0.80	0.08	**0.72**	**0.23**
Sen	0.93	0.85	0.89	0.87	0.92	0.91	0.89	0.93	0.93	0.91	0.96	0.61	**0.88**	**0.09**
MNG cases tumour core														

DSC	0.84	0.80	0.56	0.09	0.77							**0.61**	**0.31**
PPV	0.89	0.72	0.41	0.60	0.66				n.a.			**0.66**	**0.18**
Sen	0.79	0.90	0.92	0.05	0.93							**0.71**	**0.38**
LGG cases whole tumour													
DSC	0.35	0.70	0.89	0.58	0.93	0.85	0.83	0.85	0.54	0.77	n.a	**0.73**	**0.18**
PPV	0.27	0.55	0.86	0.43	0.93	0.77	0.88	0.90	0.43	0.74	n.a	**0.67**	**0.24**
Sen	0.52	0.93	0.92	0.89	0.93	0.95	0.78	0.80	0.75	0.80	n.a	**0.83**	**0.13**

The average performance results of the replicability analysis using the in-house image set and the reproducibility results are compiled in Table 4 for comparison.

Table 4: Comparison of the mean results of the reproducibility (BraTS 2015 test set) and replicability (in-house image set) analysis of DeepMedic. LGG – low grade glioma, HGG – high grade glioma, MNG – meningioma, DSC – Dice similarity coefficient, PPV – positive predictive value, Sen – sensitivity.

Data set:		In-house image set		BraTS 2015 test image set
Cases:		HGG	MNG	LGG+HGG
Tumour core	DSC	0.77	0.61	0.68
	PPV	0.72	0.66	0.83
	Sen	0.88	0.71	0.64
Cases:		LGG		LGG+HGG
Whole tumour	DSC	0.73		0.85
	PPV	0.83		0.85
	Sen	0.67		0.88

The visual evaluation of individual cases revealed a variety of causes of poor performance. In HGG visual inspection of Case #07 results showed that DeepMedic misclassified brain tissue voxels in the vicinity of the tumour core (Figure 2, top row). A similar problem was observed in Case #12 (Figure 2, middle row). The algorithm failed to segment a tumour in MNG Case #04 (Figure 2, bottom row). While the tumour location and appearance (uncharacteristic for glioma) may be the reason for a poor result, we also note that the brain mask generated in the preprocessing by BraTS Processor failed to include a part of the reference label. For LGG the algorithm achieved relatively poor results for Cases #01 and #09. The results obtained for LGG Case #01 revealed a segmentation error as a result of a preprocessing error: the brain mask included periocular tissue that was classified as tumour by the segmentation algorithm (Figure 3, top row). In LGG Case #09, DeepMedic labelled a substantial portion of the brain that was not included in the reference segmentation (Figure 3, bottom row).

Proposed updates to the checklist

From our results we deducted that insufficient description of the preprocessing was the main obstacle to reproducing Pereira's et al.[11] results. We therefore present an updated reproducibility and replicability checklist for medical segmentation studies (Table 5).

Table 5: A suggested reproducibility and replicability checklist for automatic medical image segmentation studies.
Data set – description of the image data set used for model development and validation:

- › Image acquisition parameters ☐
- › Data set size ☐

- Data excluded + reason ☐
- Link to the data set (if available) ☐

Data set preprocessing – description of the processing steps applied to the raw images before they can be fed to the segmentation model:

- List of all processing steps and corresponding parameters developed for the implementation ☐
- List of processing steps not included in the implementation (when segmentation model developed and validated on partially preprocessed data) ☐
- Statement if proprietary software was used ☐
- Link to the source code + dependencies ☐

Segmentation model – description of the model's architecture used for the segmentation:

- Description of the model (layers, nodes, functions, etc.) ☐
- Trained model ☐
- Framework used to build the model + version ☐
- Statement if proprietary software was used ☐
- Link to the source code + dependencies ☐

Postprocessing – description of all processing steps and corresponding parameters applied to the output of the segmentation algorithm before evaluation:

- List of all processing steps and corresponding parameters developed for the implementation ☐
- Statement if proprietary software was used ☐
- Link to the source code + dependencies ☐

Model development – description of the training/validation and optimization strategies:

- Augmentation transformations and corresponding parameters used for training ☐
- Training/validation/testing split ☐
- Final training sample size ☐
- CV strategy + number of folds / number of training and evaluation runs ☐
- Optimization algorithm + reference ☐
- Hyperparameter selection strategy ☐
- Hyperparameters (learning rate a, batch size n, drop-out d) ☐
- Link to the training source code + dependencies ☐

Computing infrastructure – description of the hardware used:

- Name ☐
- Class of the architecture ☐
- Memory size ☐

Model evaluation – description of the model evaluation:

- Metrics average + variations ☐
- Reference segmentation source ☐
- Training and testing runtime ☐
- Link to the evaluation source code or platform ☐

DISCUSSION

Reproducibility and replicability of scientific results are the foundation of evidence-based medicine. This work shows that current guidelines for publishing validation studies on deep-learning algorithms do not go far enough. While attempting to reproduce the two studies on MR brain lesion segmentation that were identified as meeting the current recommendations for requirements,[3] we found that only one of them was reproducible based on the published information. Remarkably, even after consultation with the authors of the second method, we were not able to obtain satisfactory segmentation results with their method.

We furthermore attempt to externally validate the findings reported for DeepMedic on a set of own data. We found that the available preprocessing pipeline is not free from producing errors, which directly influences the segmentation outcome. Moreover, we observed a poorer performance of the algorithm in MNG cases. This is, however, a somewhat expected behaviour since the training set did not contain any MNG tumours. On the other hand, visual inspection also revealed potential DeepMedic segmentation errors arising from preprocessing errors. Nonetheless, our results acquired with the BraTS-Processor and DeepMedic are promising, and we have begun to explore the potential of this pipeline for clinical application. Unfortunately, the experience gained through this study suggests that the available algorithms are not, in their present form, ready to be implemented in clinical routines. This, despite their meeting the recommended criteria for reproducibility as outlined by Pineau et al.[6, 8] and Renard, et al.[3] Improving the reproducibility of technical validation studies of DL segmentation methods will lay a foundation for producing strong evidence for what algorithms work best, when, and why. It will furthermore facilitate creating standardized evaluation frameworks and create a solid base for implementing DL tools in clinical routines.

Reproducibility criteria

The items that Renard et al.[3] identified as necessary to reproduce a DL methodology study are divided into information about hyperparameters (optimization, learning rate, drop-out, batch size) and the data set used (training proportion, data augmentation, and validation set). All these items are indeed included in the two studies we attempted to reproduce.[10, 11] The current recommendations, however, do not sufficiently stress the importance of thorough documentation of the image preprocessing chain.

The approach to preprocessing of the training and testing data is different between the two highlighted segmentation studies. The authors of DeepMedic guarantee optimal performance of the algorithm on images prepared for the BraTS segmentation challenge (skull stripping, spatial normalization, and resampling) with an additional intensity normalization step. Pereira's method, on the other hand, achieved its reported high accuracy after more complex preprocessing had been applied. For our study on Pereira's method, intensities of the whole images were corrected for field inhomogeneity, and histograms normalized across each sequence. The final preprocessing step involved patch normalization. These procedures were not explicitly described. We requested the missing information from the authors, and while they were supportive in principle, they were unable to supply the patch intensity information. To compensate, we extracted the mean and standard deviation from the training images by collecting intensity information of patches sampled from various brain regions to ensure class balance. We imposed a condition that for a given class, a certain percentage of patch pixels are labelled as that class. The values of mean and standard deviation depended on the percentage value, and we did not succeed at finding a value that would improve the segmentation results. Instead, we resorted to normalizing whole testing images to have zero mean and unit variance but the dominance of the intensities of healthy tissue skewed the estimated parameters. Unsurprisingly, the results show poor accuracy due to our inability to reproduce the intensity normalization procedures conducted in the original study.

The problem of insufficient reporting of the preprocessing procedures has been recognized previously.[5] While preprocessing may be less important in the context of segmentation challenges, evaluating the whole processing chain, from raw images to the final segmentation , is crucial in the context of application to independently collected data. Without the ability to reproduce the whole processing chain, meaningful method comparison and validation on external data becomes impossible.

Our findings prompt us to propose a significant modification to the previously reported reproducibility checklist by Pineau et al.[6] and Renard et al.'s guidelines.[3] We present this new checklist in Table 5. First, we add what we conclude to be a necessary and sufficient description of the preprocessing. Second, we regroup the items to provide a clearer distinction between the various elements and aspects that are involved in the algorithm development vs. the validation of the medical image segmentation tool: such a structure for providing a more transparent and easily implemented way of reporting is specifically designed to help those who seek to reproduce and replicate. More generally, these modifications are critical to improving the reproducibility and replicability of medical image segmentation methods.

Replication analysis

The external validation was conducted on locally acquired images. We cannot draw definitive conclusions regarding DeepMedic's performance in a clinical setting. Because of our small sample size, we also cannot make inferences about applying deep-learning methods trained on glioma cases to other tumour cases. Our results, however, are promising. The analysis further highlighted how essential the preprocessing chain is for accurate brain tumour segmentation with DeepMedic and likely with any other DL segmentation method.

In our pipeline, we used BraTS-Processor to take advantage of a tool that will automatically apply all the preprocessing steps that were also applied to the training set. Our analysis revealed segmentation errors that could be traced to errors in the preprocessing. Cases of errors in the skull stripping, which we observed in the in-house data, have been reported previously[32, 33] and will likely cause occasional problems in the future. Nonetheless, the processing pipeline generates segmentations that, even if erroneous in a few cases, will be easy to correct if the operator is equipped with a suitable interactive label editing tool. Developers of clinical tools should be aware of the issue and enable users to easily remove mislabelled regions.[34]

In addition to the noted preprocessing errors, we encountered another problem that likely influenced the results: the BraTS-Processor outputs images in the BraTS space. To evaluate the automatic segmentations quantitatively, we had to transform the reference segmentations from the native space to the BraTS space as well. This resulted in visible distortions to the reference segmentations. Accordingly, the results we presented (Table 5) likely underestimate the performance of the method (BraTS-Processor + DeepMedic) on externally acquired data. For a more accurate evaluation of a given processing pipeline, reference segmentations should be delineated on images in the BraTS space. While it may not be feasible in retrospective studies, it is a vital study design step for prospective studies. Clinical practitioners are more accustomed to the MNI152 reference space.[35] More images would become available if images could be registered to the MNI152 space instead of the BraTS space during preprocessing in the BraTS Toolkit.

CONCLUSIONS

Established reproducibility criteria for studies developing and validating DL lesion segmentation algorithms are not sufficient with regard to the preprocessing steps. The results of the reproducibility analysis led us to propose a new reproducibility checklist for medical image segmentation studies, especially if clinical utility of the algorithms is the goal. We further highlighted that even a fully reproducible preprocessing method is prone to errors on routine clinical images, which is likely to impair the segmentation outcome. We encourage researchers in the field of medical image segmentation to follow our modified checklist and assess it in terms of practical utility.

ETHICS APPROVAL

The data were acquired under approval by the Swedish Ethical Review Authority (Dnr 702-18), which waived the requirement of informed consent.

AUTHOR CONTRIBUTIONS

EG conducted the study and led the writing of the article. RAH was the main supervisor and consultant of the study progress and design choices. JS and IB-B were co-supervising the study progress at all stages. AJ and TD provided us with the in-house collected images, reference segmentations, and design input for the external validation. All co-authors collaborated on manuscript composition and editing.

FUNDING STATEMENT

This work was supported by VGR InnovationsFonden (VGRINN-940050) and ALF funds (SU2018-03591 and ALFGBG-925851).

COMPETING INTERESTS

The authors declare that they have no competing interests.

DATA SHARING

The code generated for this study is available from https://github.com/emiliagyska/repro_study.git

ACKNOWLEDGEMENTS

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. The authors would like to thank S. Pereira for providing us with details of his study, his support, and interest in our work.

REFERENCES

[1] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[2] J. E. Park, P. Kickingereder, and H. S. Kim, "Radiomics and Deep Learning from Research to Clinical Workflow: Neuro-Oncologic Imaging," *Korean J. Radiol.*, vol. 21, no. 10, pp. 1126–1137, Oct. 2020, doi: 10.3348/kjr.2019.0847.

[3] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, vol. 10, Aug. 2020, doi: 10.1038/s41598-020-69920-0.

[4] Z. C. Lipton and J. Steinhardt, "Research for practice: troubling trends in machine-learning scholarship," *Commun. ACM*, vol. 62, no. 6, pp. 45–53, May 2019, doi: 10.1145/3316774.

[5] E. Gryska, J. Schneiderman, I. Björkman-Burtscher, and R. A. Heckemann, "Automatic brain lesion segmentation on standard magnetic resonance images: a scoping review," *BMJ Open*, vol. 11, no. 1, p. e042660, Jan. 2021, doi: 10.1136/bmjopen-2020-042660.

[6] J. Pineau *et al.*, "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)," *ArXiv200312206 Cs Stat*, Apr. 2020, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/2003.12206

[7] B. Haibe-Kains *et al.*, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, Art. no. 7829, Oct. 2020, doi: 10.1038/s41586-020-2766-y.

[8] J. Pineau, "Machine Learning Reproducibility Checklist." https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (accessed Oct. 01, 2021).

[9] "MICCAI 2021 - 24. International Conference On Medical Image Computing & Computer Assisted Intervention." https://miccai2021.org/en/ (accessed Jun. 21, 2021).

[10] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017, doi: 10.1016/j.media.2016.10.004.

[11] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.

[12] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018, doi: 10.1109/ACCESS.2017.2788044.

[13] E. National Academies of Sciences, *Reproducibility and Replicability in Science*. 2019. doi: 10.17226/25303.

[14] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva, *Brain Tumor Segmentation using Convolutional Neural Networks in MRI Images*. [Online]. Available: http://dei-s2.dei.uminho.pt/pessoas/csilva/brats_cnn/

[15] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," *ArXiv Prepr.*, vol. arXiv:1210.5644, p. 9, 2012.

[16] "BRATS - SICAS Medical Image Repository." https://www.smir.ch/BRATS/Start2015 (accessed Jan. 28, 2021).

[17] T. Tieleman, G. Hinton, and others, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[18] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," *ICML*, no. 3, pp. 1139–1147, 2013.

[19] F. Bastien *et al.*, "Theano: new features and speed improvements," *ArXiv12115590 Cs*, Nov. 2012, Accessed: Jun. 14, 2021. [Online]. Available: http://arxiv.org/abs/1211.5590

[20] S. Dieleman *et al.*, "Lasagne: First release." Aug. 2015. doi: 10.5281/zenodo.27878.

[21] Data Format Working Group, "NIfTI: — Neuroimaging Informatics Technology Initiative." https://nifti.nimh.nih.gov/ (accessed Jan. 28, 2021).

[22] N. J. Tustison *et al.*, "N4ITK: improved N3 bias correction," *IEEE Trans. Med. Imaging*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.

[23] L. G. Nyul, J. K. Udupa, and Xuan Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imaging*, vol. 19, no. 2, pp. 143–150, Feb. 2000, doi: 10.1109/42.836373.

[24] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[25] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler, "The Virtual Skeleton Database: An Open Access Repository for Biomedical Research and Collaboration," *J. Med. Internet Res.*, vol. 15, no. 11, p. e245, 2013, doi: 10.2196/jmir.2930.

[26] L. Ibanez *et al.*, *The ITK Software Guide*. Kitware Inc (2018)., 2003.

[27] S. Bauer, T. Fejes, and M. Reyes, "A Skull-Stripping Filter for ITK," *Insight J.*, p. 859, Apr. 2012.

[28] F. Kofler *et al.*, "BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice," *Front. Neurosci.*, vol. 14, 2020, doi: 10.3389/fnins.2020.00125.

[29] J. C. Reinhold, B. E. Dewey, A. Carass, and J. L. Prince, "Evaluating the Impact of Intensity Normalization on MR Image Synthesis," *ArXiv181204652 Cs*, Dec. 2018, Accessed: Feb. 08, 2021. [Online]. Available: http://arxiv.org/abs/1812.04652

[30] B. B. Avants, N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee, "The Insight ToolKit image registration framework," *Front. Neuroinformatics*, vol. 8, p. 44, Apr. 2014, doi: 10.3389/fninf.2014.00044.

[31] K. Gorgolewski *et al.*, "Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python," *Front. Neuroinformatics*, vol. 0, 2011, doi: 10.3389/fninf.2011.00013.

[32] F. Kellner-Weldon *et al.*, "Comparison of perioperative automated versus manual two-dimensional tumor analysis in glioblastoma patients," *Eur. J. Radiol.*, vol. 95, pp. 75–81, Oct. 2017, doi: 10.1016/j.ejrad.2017.07.028.

[33] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F. Münte, and H. Handels, "Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences," *J. Neurosci. Methods*, vol. 240, pp. 89–100, Jan. 2015, doi: 10.1016/j.jneumeth.2014.11.011.

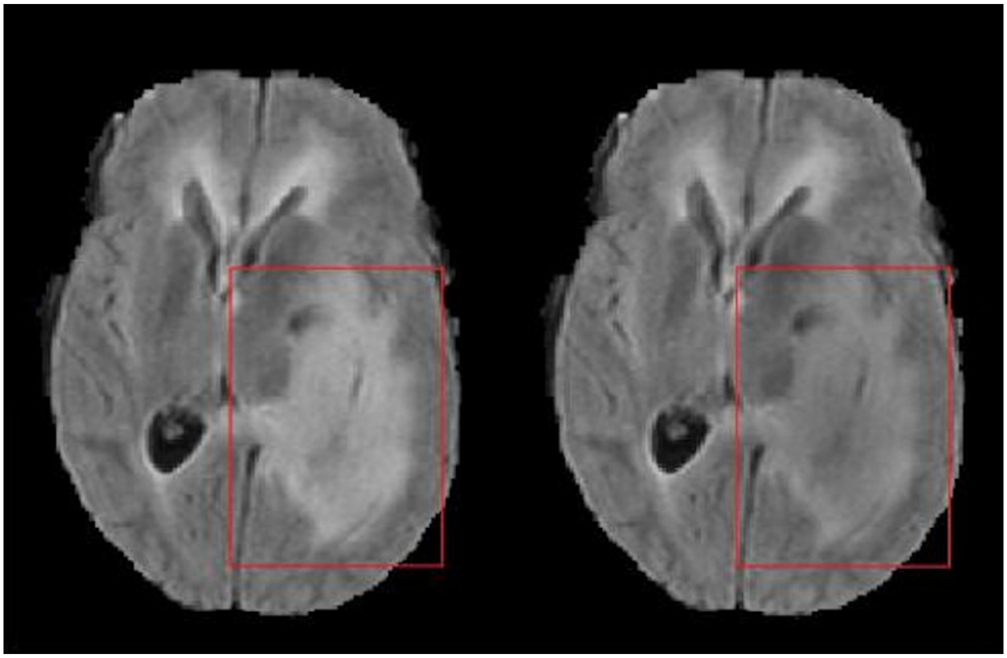
[34] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," Management Science, vol. 64, no. 3, pp. 1155–1170, Nov. 2016, doi: 10.1287/mnsc.2016.2643.

[35] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," in *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, Oct. 1993, pp. 1813–1817 vol.3. doi: 10.1109/NSSMIC.1993.373602.

Figure 1: Comparison of the field inhomogeneity correction with ANTs/Nipype (left) and SimpleITK (right). Distinct differences in the FLAIR signal intensity of tumour tissue are visible (red squares).

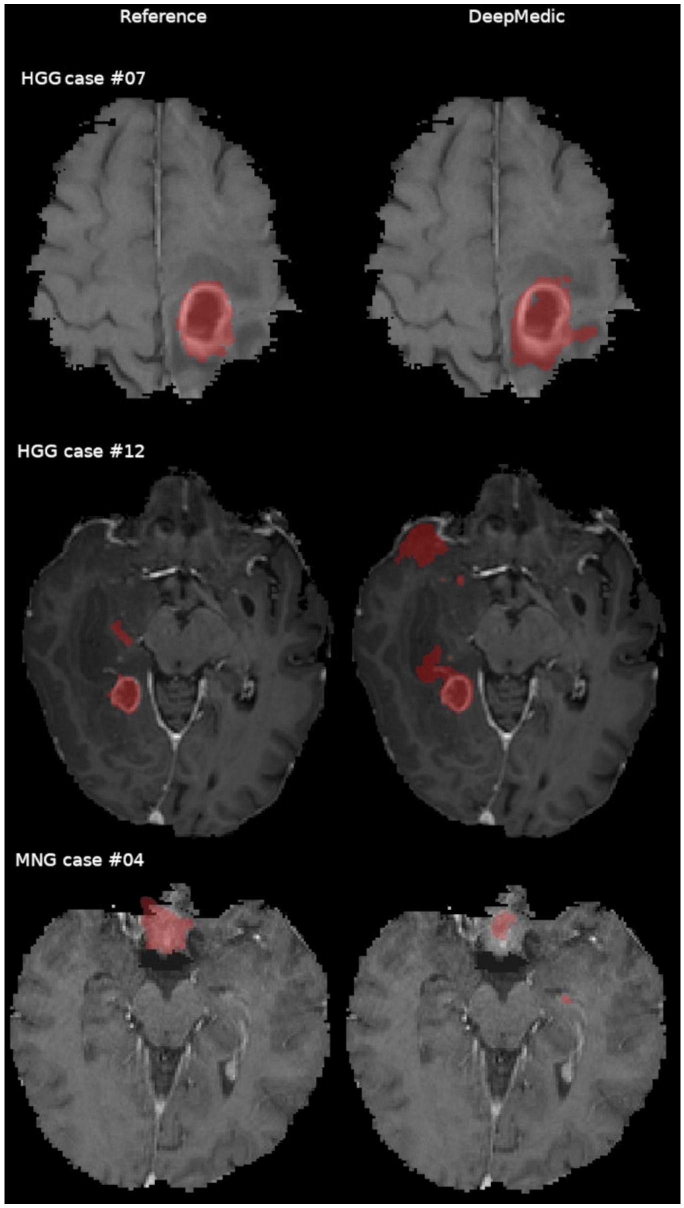
Figure 2: Comparison of the expert segmentation (reference) and DeepMedic tumour core segmentation in the in-house data for high grade glioma (HGG) and meningioma (MNG) cases overlaid on contrast enhanced T1-weighted. Voxels misclassified by DeepMedic are visible in HGG Cases #07 and #12 (top and middle row). DeepMedic failed to correctly outline the tumour and included normal brain structures in the left medial temporal lobe for meningioma Case #04 (bottom row).

Figure 3: Comparison of the expert segmentation (reference) and DeepMedic whole tumour segmentation in the in-house data for low grade glioma cases displayed overlaid on FLAIR images. Voxels misclassified by DeepMedic are visible bilaterally in the orbit in Case #01 (top row), which should have been excluded by the skull stripping procedure. In Case #09 (middle row), DeepMedic misclassified contralateral, sequence-depended FLAIR hyperintensities.



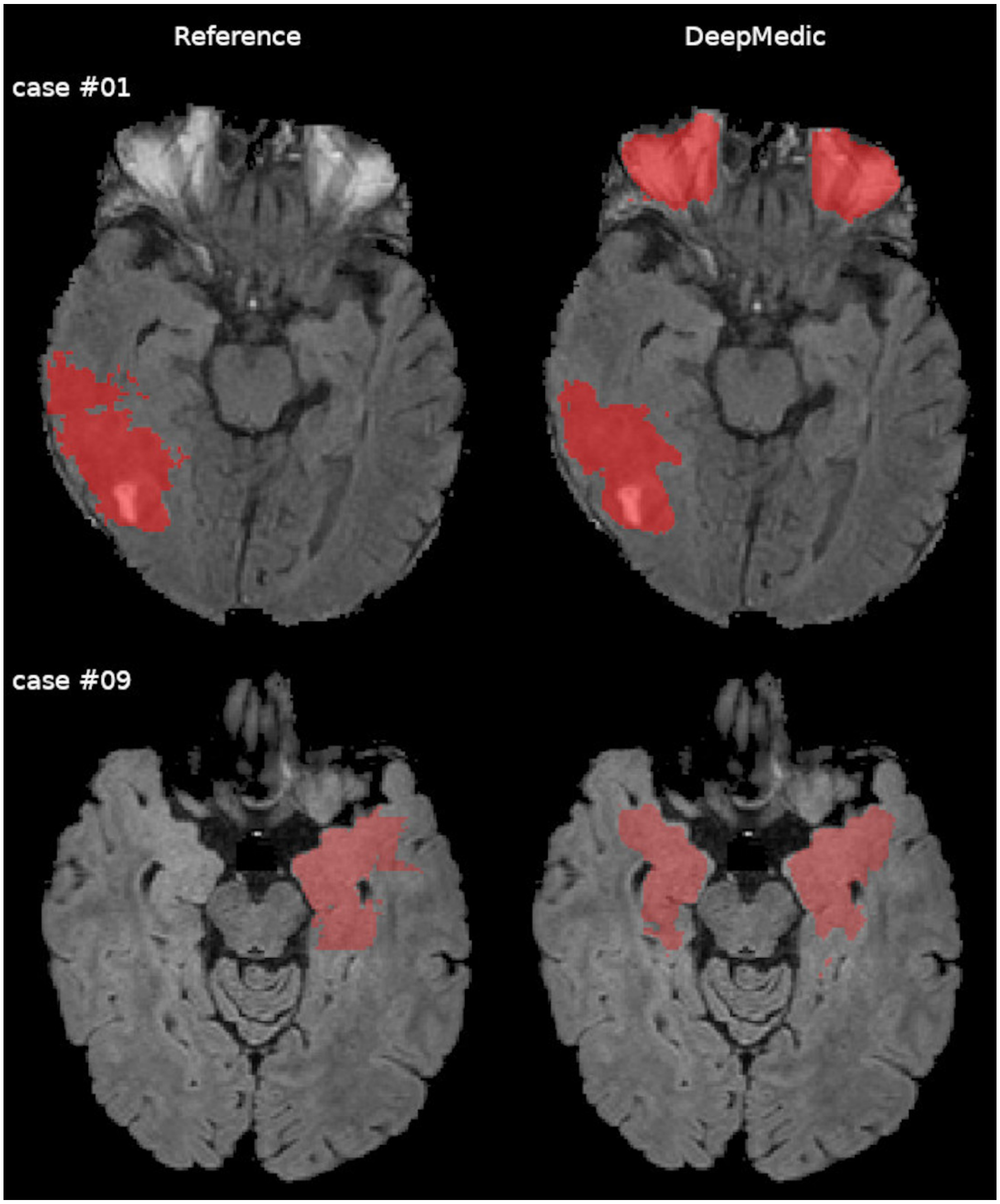
Comparison of the field inhomogeneity correction with ANTs/Nipype (left) and SimpleITK (right). Distinct differences in the FLAIR signal intensity of tumour tissue are visible (red squares).

67x44mm (300 x 300 DPI)



Comparison of the expert segmentation (reference) and DeepMedic tumour core segmentation in the in-house data for high grade glioma (HGG) and meningioma (MNG) cases overlaid on contrast enhanced T1-weighted. Voxels misclassified by DeepMedic are visible in HGG Cases #07 and #12 (top and middle row). DeepMedic failed to correctly outline the tumour and included normal brain structures in the left medial temporal lobe for meningioma Case #04 (bottom row).

94x168mm (300 x 300 DPI)



Comparison of the expert segmentation (reference) and DeepMedic whole tumour segmentation in the in-house data for low grade glioma cases displayed overlaid on FLAIR images. Voxels misclassified by DeepMedic are visible bilaterally in the orbit in Case #01 (top row), which should have been excluded by the skull stripping procedure. In Case #09 (middle row), DeepMedic misclassified contralateral, sequence-depended FLAIR hyperintensities.

93x112mm (300 x 300 DPI)

BMJ Open

Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059000.R1
Article Type:	Original research
Date Submitted by the Author:	11-May-2022
Complete List of Authors:	Gryska, Emilia; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Medical Radiation Sciences, Institute of Clinical Sciences Björkman-Burtscher, Isabella; University of Gothenburg Sahlgrenska Academy, Department of Radiology, Institute of Clinical Sciences; Sahlgrenska University Hospital, Department of Radiology Jakola, Asgeir Store; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology; Sahlgrenska University Hospital, Department of Neurosurgery Dunås, Tora; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology; University of Gothenburg Sahlgrenska Academy, Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology Schneiderman, Justin; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology Heckemann, Rolf; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Medical Radiation Sciences, Institute of Clinical Sciences
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Radiology and imaging
Keywords:	Magnetic resonance imaging < RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING, Neuroradiology < RADIOLOGY & IMAGING, STATISTICS & RESEARCH METHODS

Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study

Emilia A. Gryska[1,2*], Isabella M. Björkman-Burtscher[3,4], Asgeir S. Jakola[5,6], Tora Dunås[5,7], Justin F. Schneiderman[1,5], Rolf A. Heckemann[1,2]

1 MedTech West at Sahlgrenska University Hospital, University of Gothenburg, Gothenburg, Sweden

2 Department of Medical Radiation Sciences, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

3 Department of Radiology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

4 Department of Radiology, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden

5 Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

6 Department of Neurosurgery, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden.

7 Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Sweden

* Corresponding author: Emilia A. Gryska

Postal address: MedTech West, Röda stråket 10B, Sahlgrenska University Hospital, 413 45 Gothenburg, Sweden

E-mail address: emilia.gryska@gu.se

Telephone number: +46 720304106

Keywords: Reproducibility of Results, Deep Learning, Brain Neoplasms, Magnetic Resonance Imaging

Word count: 4 094

ABSTRACT

Objectives: To determine the reproducibility and replicability of studies that develop and validate segmentation methods for brain tumours on magnetic resonance images (MRI) and that follow established reproducibility criteria; and to evaluate whether the reporting guidelines are sufficient.

Methods: Two eligible validation studies of distinct deep learning (DL) methods were identified. We implemented the methods using published information and retraced the reported validation steps. We evaluated to what extent the description of the methods enabled reproduction of the results. We further attempted to replicate reported findings on a clinical set of images acquired at our institute consisting of high and low grade glioma (HGG, LGG), and meningioma (MNG) cases.

Results: We successfully reproduced one of the two tumour segmentation methods. Insufficient description of the preprocessing pipeline and our inability to replicate the pipeline resulted in failure to reproduce the second method. The replication of the first method showed promising results in terms of Dice similarity coefficient (DSC) and sensitivity (Sen) on HGG cases (DSC=0.77, Sen=0.88) and LGG cases (DSC=0.73, Sen=0.83), however poorer performance was observed for MNG cases (DSC=0.61, Sen=0.66). Preprocessing errors were identified that contributed to low quantitative scores in some cases.

Conclusions: Established reproducibility criteria do not sufficiently emphasize description of the preprocessing pipeline. Discrepancies in preprocessing as a result of insufficient reporting are likely to influence segmentation outcomes and hinder clinical utilization. A detailed description of the whole processing chain, including preprocessing, is thus necessary to obtain stronger evidence of the generalizability of DL-based brain tumour segmentation methods and to facilitate translation of the methods into clinical practice.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This is an independent evaluation of the reproducibility of DL-based lesion segmentation studies that follow established reporting guidelines.
- We experimentally assessed a theoretically derived reproducibility checklist for medical image segmentation studies.
- The clinical data set acquired at our institution was suitable for the replication part of the study.
- This study did not aim to enable inferences about the clinical utility of the evaluated algorithms.

INTRODUCTION

The scientific community has directed substantial efforts at developing deep-learning (DL) methods for medical image analysis. DL methods have become the default choice under the claim of superior performance to classical algorithms.[1-3] However, their outstanding performance comes at the cost of high complexity and inherent variability in model performance.[3] Consequently, assessing which model design choices determine the empirical gains is challenging.[3-5] Critics have also pointed out that scientific reporting of study designs has often been insufficient, and that the analysis of results tends to be biased towards authors' desired outcomes.[4, 6, 7] These issues present critical challenges to realizing the potential of artificial intelligence (AI) and translating promising scientific algorithms into reliable and trusted clinical decision support tools.

In our previous work,[5] we systematically explored the literature to identify whether prevalent brain lesion segmentation methods are a suitable basis for developing a tool that supports radiological brain tumour status assessment. Our findings corroborated the issues with reporting that may affect reproducibility.[5] In particular, reporting of the preprocessing steps is inadequate in many instances.

The problem has been recognized by researchers, and efforts have been made to standardize reporting practices of DL validation studies. The checklist proposed by Pineau et al.[6, 8] identifies a set of items to be reported pertaining to the presented models/algorithms, theoretical claims, data sets, code, and experimental results. The reproducibility problem in relation to the specific field of medical image segmentation was highlighted by Renard et al. in a literature review.[3] The authors present recommendations for the framework description that provides specific context for medical image segmentation. Their recommended items to be reported[3] are largely congruent with those proposed by Pineau et al.[6, 8] Renard et al.,[3] however, group their items by sources of variability in the model and evaluation framework, in contrast to grouping by scientific article section, as originally proposed by Pineau et al.[6, 8]

Furthermore, Renard et al.[3] only identified three out of twenty-nine studies included in their review to be sufficiently described according to their reproducibility recommendations. Two[9, 10] of the three were algorithms for brain tumour segmentation on magnetic resonance images (MRI). To continue our pursuit of a technically validated DL brain tumour segmentation algorithm that is suitable for clinical validation, we attempted to re-implement the two methods[9, 10].

The two DL brain tumour segmentation methods were technically validated convolutional neural networks (CNNs). Kamnitsas et al.[9] developed a 3D dual-pathway CNN with fully connected 3D conditional random fields (CRF).[11] The method will be referred to as 3D dual-path CNN in this article. The authors made the method available for independent evaluation (https://github.com/deepmedic) but did not provide a trained model. The software came with a set of configurable network parameters and requirements for the input data. The input data requirements were: images in NIfTI[12] format; images for each patient and reference labels with optional brain tissue masks (regions of interest – ROIs) had to be co-registered; all images fed to the network had to have the same voxel size; and for optimal performance, MRI signal intensities had to be standardized to have zero-mean and unit-variance within each ROI.

Pereira et al. developed a 2D single-pathway CNN, referred to as 2D single-path CNN in this article. The authors published two network architectures (HGG – high grade glioma and LGG – low grade glioma) with trained weights,[13]. The preprocessing described in the original publication consisted of bias field correction with N4ITK,[14] followed by intensity normalization[15] of each image. The input patch intensities were finally normalized with the mean and standard deviation calculated from the training patches across each sequence. A roughly similar number of patches was extracted for each class (approximately 50 000 per class for HGG to match the number of patches extracted for training as stated in the original article). The segmentation result was further processed by removing clusters of voxels smaller than a predefined threshold of 10 000 mm^3 and 3 000 mm^3 in HGG and LGG, respectively.

The aim of this study was therefore to determine the reproducibility and replicability of the two methods for brain tumour segmentation[9, 10] that Renard et al. identified as adequately reported;[3] and to evaluate whether Renard's and Pineau's reproducibility recommendations are sufficient also for the task to segment an in-house clinical data set of brain tumours.

MATERIAL AND METHODS

Overview

The study design is based on the assumption that the reproducibility items proposed by Renard et al. are sufficient for reproduction and replication. We used the definitions of reproduction and replication from the National Academies of Sciences, Engineering and Medicine,[16] which Pineau et al. also refer to.[6] Renard et al. identified two methods for brain lesion segmentation[9, 10] as adequately reported,[3] and we chose these two for the present study. Our goal was to implement the respective original methods with all processing steps and parameters and test them on the same data on which they were originally validated (reproducibility). As a measure of success, we compared quantitative results on segmentation accuracy to those reported in the original studies. We then attempted to replicate[6, 16] the findings: we performed an external validation on a clinically obtained data set from our institution.

Patient and Public Involvement

No patient involved.

Statistical analysis

We provide descriptive statistics (means, and when possible standard deviations) of segmentation evaluation metrics. The metrics we used are: Dice similarity coefficient – DSC, positive predictive value – PPV, and sensitivity.

Reproducibility analysis

Evaluated segmentation algorithms

We implemented the two previously proposed DL algorithms for brain tumour segmentation: 3D dual-path CNN[9] and 2D single-path CNN.[10] In Table 1 these algorithms are described in compliance with the reproducibility categories listed by Renard et al.,[3] together with libraries and computational parameters we used in our implementations. For our implementation, we used hyperparameters reported in the original articles. We trained the 3D dual-path CNN and tested both algorithms on a cluster with a Tesla V100 GPU (5120 cores; Nvidia Corp., Santa Clara, CA, USA), 32 GB RAM, and two 8-core Xeon Gold 6244 @ 3.60GHz processors (Intel Corp., Santa Clara, CA, USA).

Table 1: Description of the two algorithms implemented in the reproducibility analysis, 3D dual-path CNN[9] and 2D single-path CNN,[10] according to the reproducibility categories proposed by Renard et al.[3] All the parameters and versions found in the first part of the table were specified in the original articles. In the part **"Our implementation middleware"**, we specify the Python version and libraries used for our implementations. CNN – convolutional neural networks, CRF – conditional random field, CV – cross-validation, DSC – Dice similarity coefficient, FC – fully connected, HGG – high grade glioma, LGG – low grade glioma.

Main category	Sub-category	3D dual-path CNN	2D single-path CNN
Algorithm/model	Description of the DL architecture	Dual-path 3D CNN with a fully connected 3D CRF.[11]	Single-path 2D CNN; two network architectures for HGG and LGG.
Dataset description	Image acquisition parameters Image size Data set size Link to the data set	BraTS 2015 dataset[17]	
Preprocessing description	Data excluded + reason	none	none
	Augmentation transformation	Sagittal reflection of images	Rotation with multiples of 90° angles
	Final sample size	Not specified	~1 800 000 for HGG ~1 340 000 for LGG
Training/validation/ testing split	Explanation if validation set not created	Training and testing sets provided by the BraTS challenge	
CV strategy + number of folds	Not specified	5-fold CV on training set (n=274)	1 subject in both HGG (n=220) and LGG (n=54)
Optimization strategy	Optimization algorithm + reference	RMSProp optimizer[18] and Nesterov's momentum[19]	Stochastic Gradient Descent and Nesterov's momentum[19]
	Hyperparameters (learning rate *a*, batch size *n*, drop-out *d*)	$a = 10^{-3}$ (halved when the convergence plateaus); $n = 10$ $d = 50\%$ (in the last 2 hidden layers)	$a_{inital} = 0.003$ $a_{final} = 0.00003$ $n = 128$ $dHGG = 0.1$ (in FC layers) $dHGG = 0.5$ (in FC layers)
	Hyperparameter selection strategy	CRF: 5-fold CV on a training subset HGG (n=44) and LGG (n=18)	Validation using 1 subject in both HGG (n=220) and LGG (n=54)
Computing infrastructure	Name, class of the architecture, and memory size	NVIDIA GTX Titan X GPU using cuDNN v5.0, 12GB	GPU NVIDIA GeForce GTX 980
Middleware	Toolbox used/in-house code + build version	Theano[20] Python 3.6.5, Tensorflow 2.0.0/1.15.0, Nibabel 3.0.2 Numpy 1.18.2	Theano 0.7.0[20] Lasagne 0.1dev[21] Python 2.7.10 Numpy 1.9.2
	Source code link + dependencies	https://github.com/deepmedic	http://dei-s2.dei.uminho.pt/pessoas/csilva/brats_cnn/
Evaluation	Metrics average + variations	Mean of DSC, Precision, and Sensitivity (calculated by the online evaluation system)	Boxplot and mean of DSC (calculated by the online evaluation system)

Our implementation middleware

Python version	3.8.2	3.7.4
DL library	Tensorflow 2.2.1	Theano (git version eb6a412), Lasagne (git version 5d3c63c)
Numpy	1.18.5	1.17.3
Nibabel	3.0.2	3.2.1

Image data set used for reproducibility analysis

Both algorithms were originally validated in the 2015 Brain Tumor Segmentation Challenge (BraTS),[22] which consists of training and testing image sets of patients diagnosed with HGG and LGG. The training set contains 274 examinations (HGG n = 220, LGG n = 54). Each examination consists of T1-weighted (T1w) images before and after injection of contrast material (CM), T2w, and FLAIR (fluid-attenuated inversion recovery) images. The training data set additionally contains manual segmentations of tumour structures that serve as a criterion standard and delineate necrotic core, contrast-enhancing (CE) core, non-CE core, and oedema. For the test set containing 110 examinations the criterion standard segmentations are not publicly available. Users can upload their segmentation results to an online system[17, 23] that internally compares the results with the hidden reference to determine per-case metrics (DSC, PPV, sensitivity, and kappa). The system then returns summary measures (means and ranking position) to the user. Images in both sets are provided in .mha format and have been preprocessed with spatial normalization,[24] skull-stripping,[25] and resampling to an isotropic resolution of 1 mm^3 (linear interpolator).

Outcome parameters

We experimentally evaluated whether the two methods that Renard et al. identified as reproducible according to their proposed criteria[3] were possible to reproduce. Specifically, we examined whether enough information was given in the original articles or supplementary information for each processing step. If re-implementation did not reproduce the originally reported results, we contacted the authors directly to follow up on any missing details and added this information to the results. Pereira et al. supplied a pre-trained model;[13] for 3D dual-path CNN, we trained our re-implementation on the BraTS 2015 training data. Thereafter, we segmented the BraTS 2015 test set with both methods. We submitted the resulting segmentations to the online evaluation system[17] and recorded the summary measures returned (mean DSC, mean sensitivity, and mean PPV). Finally, we compared the summary measures with those available in the original publications.

Replication analysis

Evaluated segmentation algorithm

Only the 3D dual-path CNN was successfully re-implemented (cf. Results – Reproducibility study). External validation (replication analysis) on in-house clinical data was therefore carried out with this method. The segmentation models trained on the BraTS training data in the reproducibility analysis were applied to our dataset using a workstation with an Intel Core i7-6700HQ CPU @ 2.60 GHz processor and Nvidia GTX960M graphics card.

Image data set used for the replication analysis

The clinical in-house testing data set consisted of images from 27 cases (HGG n = 12; LGG n = 10; meningioma – MNG n = 5). The set was selected for this study from a larger sample of image data. Data were anonymized and inclusion criteria were pre-operative examinations, availability of manual expert reference segmentations, and imaging findings typical for the included types of pathology.

As in the BraTS data set, each MR examination included non-CM T1w, CM T1w, T2w, and FLAIR images. The images were provided in NIfTI[12] format. Since we used a model trained on BraTS data to segment these images, we used the BraTS-Processor module from the BraTS Toolkit[26] for preprocessing. Binary lesion segmentations had been prepared by trained personnel and revised by a senior neurosurgeon (AJ). Whole-tumour labels generated by delineation of T2/FLAIR hyperintensities were used for LGG. For HGG and MNG, the tumour core label was used, which had been delineated on CM T1w images and included CE tumour as well as any components enclosed by CE tumour. The reference segmentations were registered from the native space to the BraTS space following the transformation steps and using the registration matrices generated by the BraTS-Processor.[26]

Outcome parameters

The replicability of the 3D dual-path CNN was assessed by comparing DSC, sensitivity, and PPV derived from processing the clinical in-house data with those provided by the online system[17] during the reproducibility analysis on the BraTS test set. We visually evaluated individual cases to determine causes of segmentation errors.

Based on findings from the reproducibility and the replication analysis we reviewed recommendations on reporting items proposed by Renard et al.[3] and Pineau et al.[8] Challenges and failures in our attempts at reproduction and replication were documented and examined throughout the processes above. We then assessed and summarized these outcomes with suggested specific improvements to the reproducibility items for lesion segmentation on magnetic resonance images for brain segmentation.

RESULTS

Reproducibility study

3D dual-path CNN

BraTS data fulfilled most of the input requirements for the 3D dual-path CNN, apart from the format and the image intensity normalization. To reproduce the study, all images were converted to NIfTI format, and MR signal intensities were normalized to have zero-mean and unit-variance within each ROI. We implemented these steps using SimpleITK for image conversion and an in-house python program for signal intensity normalization. Since the BraTS images are already skull-stripped, we generated brain masks for each patient by thresholding each image to include only non-zero voxels in order to reduce the runtime of the algorithm. The only changes we made in the 3D dual-path CNN configuration file were to set the number of input channels to all four available, as described in the original article (default in the source code was CE T1w and FLAIR), and to specify not to perform validation of the available samples, as the hyperparameters had already been defined for the model. Training the algorithm took approximately 27 hours, and testing took 14.5 minutes.

The quantitative evaluation shows that our re-implementation and testing of the 3D dual-path CNN on the BraTS 2015 data set achieved comparable results to those presented in the original study (Table 2). We therefore deem the method reproducible.

Table 2: Reproducibility results on BraTS 2015 presented in the original paper for the 3D dual-path CNN[9] and for the 2D single-path CNN[10] (original) and for our independent reproducibility analysis (this work). Our analysis was carried out for high grade glioma (HGG) and low grade glioma (LGG) model parameters of the 2D single-path CNN. The results were congruent with the original analysis for the 3D dual-path CNN but they show an unsuccessful attempt to reproduce the 2D single-path CNN validation. The higher score in each column is emphasized in bold. Measures of dispersion or significance of differences were not available for the original method evaluation. CE – contrast-enhanced.

	Dice similarity coefficient			**Positive predictive value**			**Sensitivity**		
	Whole tumour	**Tumour core**	**CE tumour**	**Whole tumour**	**Tumour core**	**CE tumour**	**Whole tumour**	**Tumour core**	**CE tumour**
3D dual-path CNN									
Original	0.85	0.67	0.63	0.85	**0.85**	**0.63**	0.88	0.61	0.66
This work	0.85	**0.68**	**0.64**	0.85	0.83	0.62	0.88	**0.64**	**0.70**
2D single-path CNN									
Original	**0.78**	**0.65**	**0.75**	-	-	-	-	-	-
This work (HGG)	0.36	0.25	0.17	0.36	0.21	0.29	0.54	0.58	0.17

This work (LGG)	0.25	0.14	0.13	0.40	0.51	0.37	0.25	0.10	0.10

2D single-path CNN

The preprocessing description by Pereira et al. lacked certain parameters pertaining to the intensity normalization: percentile points used to create a reference histogram for each sequence and glioma grade, and intensity parameters of the training patches. Furthermore, it was not specified which model architecture was used on the BraTS 2015 test set, where the data include both HGG and LGG. Despite the missing parameters, we made an attempt to reproduce the study. We used N4ITK bias field correction (as implemented in SimpleITK) with default parameters and a histogram normalization procedure adapted from Reinhold et al.[27] We decided on this implementation instead of the corresponding function in SimpleITK, because the latter requires a reference image or histogram, neither of which was available. For the final patch-normalization step, the intensity parameters were not available, so we normalized each test image ROI to have zero-mean and unit-variance. Finally, the results were post-processed according to the procedure described by the authors. The testing time of the 2D single-path CNN was approximately 8 hours.

As the attempt was unsuccessful (results of the quantitative evaluation presented in Table 2), we approached the lead author of the method and requested the missing information. The author generously provided information on the bias field correction as well as image histogram normalization parameters.

Following this input, the N4ITK bias field correction was conducted using the implementation in ANTs[28] with the wrapper in Nipype[29] with the following parameters specified: *n_iterations* = [20, 20, 20, 10], *dimension* = 3, *bspline_fitting_distance* = 200, *shrink_factor* = 2, *convergence_threshold* = 0. A visual inspection of the field inhomogeneity correction with ANTs/Nipype and the parameters given versus SimpleITK showed signal intensity differences in the tumour region (Figure 1) that plausibly explained the failure to reproduce.

The implementation of Nyul's algorithm[15] for intensity normalization was developed in the lead author's former lab, and the author was not at liberty to share the code. Instead, the author provided percentile points and corresponding intensity landmarks for each MR sequence used in their implementation. In the original study, however, the authors trained separate sets of parameters for LGG and HGG and could not retrieve the patch intensity parameters for patch normalization. To compensate, we extracted the mean and standard deviation from the training images by collecting intensity information of patches sampled from various brain regions to ensure class balance. We imposed a condition that for a given class, a certain percentage of patch pixels are labelled as that class. The values of mean and standard deviation depended on the percentage value, and we did not succeed at finding a value that would improve the segmentation results. At this point, we decided not to pursue further efforts to reproduce the study.

Replication analysis

The replication analysis was conducted on the 3D dual-path CNN only. Quantitative results of the comparison of automatic segmented MR images collected in-house and expert delineations of the chosen tumour labels are presented in Table 3.

Table 3: 3D dual-path CNN[9] replication analysis results on in-house data for high grade glioma (HGG) cases and meningioma (MNG) cases evaluated on the tumour core and for low grade glioma (LGG) cases evaluated on the whole tumour label. DSC – Dice similarity coefficient, PPV – positive predictive value, Sen – sensitivity, Std. – standard deviation.

ID	01	02	03	04	05	06	07	08	09	10	11	12	Mean	Std.
HGG cases tumour core														
DSC	0.88	0.85	0.80	0.85	0.89	0.85	0.57	0.89	0.86	0.81	0.87	0.14	**0.77**	**0.22**
PPV	0.84	0.86	0.72	0.84	0.85	0.79	0.41	0.85	0.80	0.73	0.80	0.08	**0.72**	**0.23**
Sen	0.93	0.85	0.89	0.87	0.92	0.91	0.89	0.93	0.93	0.91	0.96	0.61	**0.88**	**0.09**
MNG cases tumour core														

DSC	0.84	0.80	0.56	0.09	0.77							**0.61**	**0.31**
PPV	0.89	0.72	0.41	0.60	0.66				n.a.			**0.66**	**0.18**
Sen	0.79	0.90	0.92	0.05	0.93							**0.71**	**0.38**
LGG cases whole tumour													
DSC	0.35	0.70	0.89	0.58	0.93	0.85	0.83	0.85	0.54	0.77	n.a	**0.73**	**0.18**
PPV	0.27	0.55	0.86	0.43	0.93	0.77	0.88	0.90	0.43	0.74	n.a	**0.67**	**0.24**
Sen	0.52	0.93	0.92	0.89	0.93	0.95	0.78	0.80	0.75	0.80	n.a	**0.83**	**0.13**

The average performance results of the replicability analysis using the in-house image set and the reproducibility results are compiled in Table 4 for comparison.

Table 4: Comparison of the mean results of the reproducibility (BraTS 2015 test set) and replicability (in-house image set) analysis of the 3D dual-path CNN.[9] LGG – low grade glioma, HGG – high grade glioma, MNG – meningioma, DSC – Dice similarity coefficient, PPV – positive predictive value, Sen – sensitivity.

Data set:		In-house image set		BraTS 2015 test image set
Cases:		HGG	MNG	LGG+HGG
Tumour core	DSC	0.77	0.61	0.68
	PPV	0.72	0.66	0.83
	Sen	0.88	0.71	0.64
Cases:		LGG		LGG+HGG
Whole tumour	DSC	0.73		0.85
	PPV	0.83		0.85
	Sen	0.67		0.88

The visual evaluation of individual cases revealed a variety of causes of poor performance. In HGG visual inspection of Case #07 results showed that the 3D dual-path CNN misclassified brain tissue voxels in the vicinity of the tumour core (Figure 2, top row). A similar problem was observed in Case #12 (Figure 2, middle row). The algorithm failed to segment a tumour in MNG Case #04 (Figure 2, bottom row). While the tumour location and appearance (uncharacteristic for glioma) may be the reason for a poor result, we also note that the brain mask generated in the preprocessing by BraTS Processor failed to include a part of the reference label. For LGG the algorithm achieved relatively poor results for Cases #01 and #09. The results obtained for LGG Case #01 revealed a segmentation error as a result of a preprocessing error: the brain mask included periocular tissue that was classified as tumour by the segmentation algorithm (Figure 3, top row). In LGG Case #09, the 3D dual-path CNN labelled a substantial portion of the brain that was not included in the reference segmentation (Figure 3, bottom row).

Proposed updates to the checklist

From our results we deducted that insufficient description of the preprocessing was the main obstacle to reproducing Pereira's et al.[10] results. We therefore present an updated reproducibility and replicability checklist for medical segmentation studies (Table 5).

Table 5: A suggested reproducibility and replicability checklist for automatic medical image segmentation studies. The update from the established checklists[3, 8] includes a new category **Data set preprocessing**, and a new item in Model evaluation category: **Failed cases: number and reasons.** We also regrouped the items into categories that provide a clearer structure for reporting in particular of reproducibility and replicability studies.

Data set – description of the image data set used for model development and validation:

- Image acquisition parameters ☐
- Data set size ☐
- Data excluded + reason ☐
- Link to the data set (if available) ☐

Data set preprocessing – description of the processing steps applied to the raw images before they can be fed to the segmentation model:

- List of all processing steps and corresponding parameters developed for the implementation ☐
- List of processing steps not included in the implementation (when segmentation model developed and validated on partially preprocessed data) ☐
- Statement if proprietary software was used ☐
- Link to the source code + dependencies ☐

Segmentation model – description of the model's architecture used for the segmentation:

- Description of the model (layers, nodes, functions, etc.) ☐
- Trained model ☐
- Framework used to build the model + version ☐
- Statement if proprietary software was used ☐
- Link to the source code + dependencies ☐

Postprocessing – description of all processing steps and corresponding parameters applied to the output of the segmentation algorithm before evaluation:

- List of all processing steps and corresponding parameters developed for the implementation ☐
- Statement if proprietary software was used ☐
- Link to the source code + dependencies ☐

Model development – description of the training/validation and optimization strategies:

- Augmentation transformations and corresponding parameters used for training ☐
- Training/validation/testing split ☐
- Final training sample size ☐
- CV strategy + number of folds / number of training and evaluation runs ☐
- Optimization algorithm + reference ☐
- Hyperparameter selection strategy ☐
- Hyperparameters (learning rate *a*, batch size *n*, drop-out *d*) ☐
- Link to the training source code + dependencies ☐

Computing infrastructure – description of the hardware used:

- Name ☐
- Class of the architecture ☐
- Memory size ☐

Model evaluation – description of the model evaluation:

- Metrics average + variations ☐
- Reference segmentation source ☐
- Failed cases: number and reasons ☐
- Training and testing runtime ☐
- Link to the evaluation source code or platform

DISCUSSION

Reproducibility and replicability of scientific results are the foundation of evidence-based medicine. In this work we show that current guidelines for publishing validation studies on deep-learning algorithms are incomplete. While attempting to reproduce the two studies on MR brain lesion segmentation that were identified as meeting current reproducibility recommendations,[3] we found that only one of them was reproducible based on the published information. Remarkably, even after consultation with the authors of the second method, we were not able to obtain satisfactory segmentation results with their method. Our claims of reproducibility / non-reproducibility could not be supported with advanced statistical analysis; the online evaluation system[17] (used to evaluate the segmentations in the original validation papers and our study) provides arithmetic means of the evaluation metrics without measures of dispersion. The small sample size of the in-house data along with the difference in tumour components segmented as a reference for HGG (tumour core) and LGG (whole tumour) further precludes a meaningful analysis of the statistical difference between the results obtained in the reproducibility and replicability analysis. We believe that our findings are nevertheless sufficient to support our conclusions.

We furthermore attempt to externally validate the findings reported for the 3D dual-path CNN on a set of own data. We found that the available preprocessing pipeline is not free from producing errors, which directly influences the segmentation outcome. Moreover, we observed a poorer performance of the algorithm in MNG cases. This is, however, a somewhat expected behaviour since the training set did not contain any MNG tumours. On the other hand, visual inspection also revealed potential the 3D dual-path CNN segmentation errors arising from preprocessing errors. Nonetheless, our results acquired with the BraTS-Processor and the 3D dual-path CNN are promising, and we have begun to explore the potential of this pipeline for clinical application. Unfortunately, the experience gained through this study suggests that the available algorithms are not, in their present form, ready to be implemented in clinical routines. This, despite their meeting the recommended criteria for reproducibility as outlined by Pineau et al.[6, 8] and Renard, et al.[3] Improving the reproducibility of technical validation studies of DL segmentation methods will lay a foundation for producing strong evidence for what algorithms work best, when, and why. It will furthermore facilitate creating standardized evaluation frameworks and create a solid base for implementing DL tools in clinical routines.

Reproducibility criteria

The items that Renard et al.[3] identified as necessary to reproduce a DL methodology study are divided into information about hyperparameters (optimization, learning rate, drop-out, batch size) and the data set used (training proportion, data augmentation, and validation set). All these items are indeed included in the two studies we attempted to reproduce.[9, 10] The current recommendations, however, do not sufficiently stress the importance of thorough documentation of the image preprocessing chain.

The approach to preprocessing of the training and testing data is different between the two highlighted segmentation studies. The authors of the 3D dual-path CNN guarantee optimal performance of the algorithm on images prepared for the BraTS segmentation challenge (skull stripping, spatial normalization, and resampling) with an additional intensity normalization step. The 2D single-path CNN, on the other hand, achieved its reported high accuracy after more complex preprocessing had been applied. For our study on the, intensities of the whole images were corrected for field inhomogeneity, and histograms normalized across each sequence. The final preprocessing step involved patch normalization. These procedures were not explicitly described. We requested the missing information from the authors, and while they were supportive in principle, they were unable to supply the patch intensity information. Unsurprisingly, the results show poor accuracy due to our inability to reproduce the intensity normalization procedures conducted in the original study.

The problem of insufficient reporting of the preprocessing procedures has been recognized previously.[5] While preprocessing may be less important in the context of segmentation challenges, evaluating the whole processing chain, from raw images to the final segmentation, is crucial in the context of application to independently collected data. Without the ability to reproduce the whole processing chain, meaningful method comparison and validation on external data becomes impossible.

Our findings prompt us to propose a significant modification to the previously reported reproducibility checklist by Pineau et al.[6] and Renard et al.'s guidelines.[3] We present this new checklist in Table 5. First, we add what we conclude to be a necessary and sufficient description of the preprocessing. Second, we regroup the items to provide a clearer distinction between the various elements and aspects that are involved in the algorithm development vs. the validation of the medical image segmentation tool: such a structure for providing a more transparent and easily implemented way of reporting is specifically designed to help those who seek to reproduce and replicate. More generally, these modifications are critical to improving the reproducibility and replicability of medical image segmentation methods. Since our updates are based on reproducibility and replicability of only two segmentation algorithms, we encourage researchers to comprehensively evaluate our checklist by including a broader selection of independently implemented algorithms for medical image segmentation.

Replication analysis

The external validation was conducted on locally acquired images. We cannot draw definitive conclusions regarding the 3D dual-path CNN's performance in a clinical setting as statistical analysis would not be meaningful; in the in-house data, we evaluated separately tumour core label in high grade glioma (HGG) examinations and whole tumour label in low grade glioma (LGG) examinations. The BraTS evaluations for both tumour components are, on the other hand, done on a mix of HGG and LGG cases. Because of our small sample size, we also cannot make inferences about applying deep-learning methods trained on glioma cases to other tumour cases. Our results, however, are promising. The analysis further highlighted how essential the preprocessing chain is for accurate brain tumour segmentation with the 3D dual-path CNN and likely with any other DL segmentation method.

In our pipeline, we used BraTS-Processor to take advantage of a tool that will automatically apply all the preprocessing steps that were also applied to the training set. Our analysis revealed segmentation errors that could be traced to errors in the preprocessing. Cases of errors in the skull stripping, which we observed in the in-house data, have been reported previously[30, 31] and will likely cause occasional problems in the future. Nonetheless, the processing pipeline generates segmentations that, even if erroneous in a few cases, will be easy to correct if the operator is equipped with a suitable interactive label editing tool. Developers of clinical tools should be aware of the issue and enable users to easily remove mislabelled regions.[32]

In addition to the noted preprocessing errors, we encountered another problem that likely influenced the results: the BraTS-Processor outputs images in the BraTS (MNI152[33]) space. To evaluate the automatic segmentations quantitatively, we had to transform the reference segmentations from the native space to the BraTS space as well. This resulted in visible distortions to the reference segmentations. Accordingly, the results we presented (Table 5) likely underestimate the performance of the method (BraTS-Processor + the 3D dual-path CNN) on externally acquired data. For a more accurate evaluation of a given processing pipeline, reference segmentations should be delineated on images in the BraTS space. While it may not be feasible in retrospective studies, it is a vital study design step for prospective studies.

CONCLUSIONS

Established reproducibility criteria for studies developing and validating DL lesion segmentation algorithms are not sufficient with regard to the preprocessing steps. The results of the reproducibility analysis led us to propose a new reproducibility checklist for medical image segmentation studies, especially if clinical utility of the algorithms is the goal. We further highlighted that even a fully reproducible preprocessing method is prone to errors on routine clinical images, which is likely to impair the segmentation outcome. We encourage researchers in the field of medical image segmentation to follow our modified checklist and assess it in terms of practical utility.

ETHICS APPROVAL

The data for the replication analysis were acquired under approval by the Swedish Ethical Review Authority (Dnr 702-18), which waived the requirement of informed consent.

AUTHOR CONTRIBUTIONS

EG conducted the study and led the writing of the article. RAH was the main supervisor and consultant of the study progress and design choices. JS and IB-B were co-supervising the study progress at all stages. AJ and TD provided us with the in-house collected images, reference segmentations, and design input for the external validation. All co-authors collaborated on manuscript composition and editing.

FUNDING STATEMENT

This work was supported by VGR InnovationsFonden (VGRINN-940050) and ALF funds (SU2018-03591 and ALFGBG-925851).

COMPETING INTERESTS

The authors declare that they have no competing interests.

DATA SHARING

The code generated for this study is available from https://github.com/emiliagyska/repro_study.git

ACKNOWLEDGEMENTS

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE), partially funded by the Swedish Research

Council through grant agreement no. 2018-05973. The authors would like to thank S. Pereira for providing us with details of his study, his support, and interest in our work.

REFERENCES

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[2] J. E. Park, P. Kickingereder, and H. S. Kim, "Radiomics and Deep Learning from Research to Clinical Workflow: Neuro-Oncologic Imaging," *Korean J. Radiol.*, vol. 21, no. 10, pp. 1126–1137, Oct. 2020, doi: 10.3348/kjr.2019.0847.

[3] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, vol. 10, Aug. 2020, doi: 10.1038/s41598-020-69920-0.

[4] Z. C. Lipton and J. Steinhardt, "Research for practice: troubling trends in machine-learning scholarship," *Commun. ACM*, vol. 62, no. 6, pp. 45–53, May 2019, doi: 10.1145/3316774.

[5] E. Gryska, J. Schneiderman, I. Björkman-Burtscher, and R. A. Heckemann, "Automatic brain lesion segmentation on standard magnetic resonance images: a scoping review," *BMJ Open*, vol. 11, no. 1, p. e042660, Jan. 2021, doi: 10.1136/bmjopen-2020-042660.

[6] J. Pineau *et al.*, "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)," *ArXiv200312206 Cs Stat*, Apr. 2020, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/2003.12206

[7] B. Haibe-Kains *et al.*, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, Art. no. 7829, Oct. 2020, doi: 10.1038/s41586-020-2766-y.

[8] J. Pineau, "Machine Learning Reproducibility Checklist." https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (accessed Oct. 01, 2021).

[9] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017, doi: 10.1016/j.media.2016.10.004.

[10] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.

[11] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," *ArXiv Prepr.*, vol. arXiv:1210.5644, p. 9, 2012.

[12] Data Format Working Group, "NIfTI: — Neuroimaging Informatics Technology Initiative." https://nifti.nimh.nih.gov/ (accessed Jan. 28, 2021).

[13] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva, *Brain Tumor Segmentation using Convolutional Neural Networks in MRI Images*. [Online]. Available: http://dei-s2.dei.uminho.pt/pessoas/csilva/brats_cnn/

[14] N. J. Tustison *et al.*, "N4ITK: improved N3 bias correction," *IEEE Trans. Med. Imaging*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.

[15] L. G. Nyul, J. K. Udupa, and Xuan Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imaging*, vol. 19, no. 2, pp. 143–150, Feb. 2000, doi: 10.1109/42.836373.

[16] E. National Academies of Sciences, *Reproducibility and Replicability in Science*. 2019. doi: 10.17226/25303.

[17] "BRATS - SICAS Medical Image Repository." https://www.smir.ch/BRATS/Start2015 (accessed Jan. 28, 2021).

[18] T. Tieleman, G. Hinton, and others, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[19] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," *ICML*, no. 3, pp. 1139–1147, 2013.

[20] F. Bastien *et al.*, "Theano: new features and speed improvements," *ArXiv12115590 Cs*, Nov. 2012, Accessed: Jun. 14, 2021. [Online]. Available: http://arxiv.org/abs/1211.5590

[21] S. Dieleman *et al.*, "Lasagne: First release." Aug. 2015. doi: 10.5281/zenodo.27878.

[22] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[23] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler, "The Virtual Skeleton Database: An Open Access Repository for Biomedical Research and Collaboration," *J. Med. Internet Res.*, vol. 15, no. 11, p. e245, 2013, doi: 10.2196/jmir.2930.

[24] L. Ibanez *et al.*, *The ITK Software Guide*. Kitware Inc (2018)., 2003.

[25] S. Bauer, T. Fejes, and M. Reyes, "A Skull-Stripping Filter for ITK," *Insight J.*, p. 859, Apr. 2012.

[26] F. Kofler *et al.*, "BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice," *Front. Neurosci.*, vol. 14, 2020, doi: 10.3389/fnins.2020.00125.

[27] J. C. Reinhold, B. E. Dewey, A. Carass, and J. L. Prince, "Evaluating the Impact of Intensity Normalization on MR Image Synthesis," *Proc. SPIE-- Int. Soc. Opt. Eng.*, vol. 10949, p. 109493H, Mar. 2019, doi: 10.1117/12.2513089.

[28] B. B. Avants, N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee, "The Insight ToolKit image registration framework," *Front. Neuroinformatics*, vol. 8, p. 44, Apr. 2014, doi: 10.3389/fninf.2014.00044.

[29] K. Gorgolewski *et al.*, "Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python," *Front. Neuroinformatics*, vol. 0, 2011, doi: 10.3389/fninf.2011.00013.

[30] F. Kellner-Weldon *et al.*, "Comparison of perioperative automated versus manual two-dimensional tumor analysis in glioblastoma patients," *Eur. J. Radiol.*, vol. 95, pp. 75–81, Oct. 2017, doi: 10.1016/j.ejrad.2017.07.028.

[31] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F. Münte, and H. Handels, "Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences," *J. Neurosci. Methods*, vol. 240, pp. 89–100, Jan. 2015, doi: 10.1016/j.jneumeth.2014.11.011.

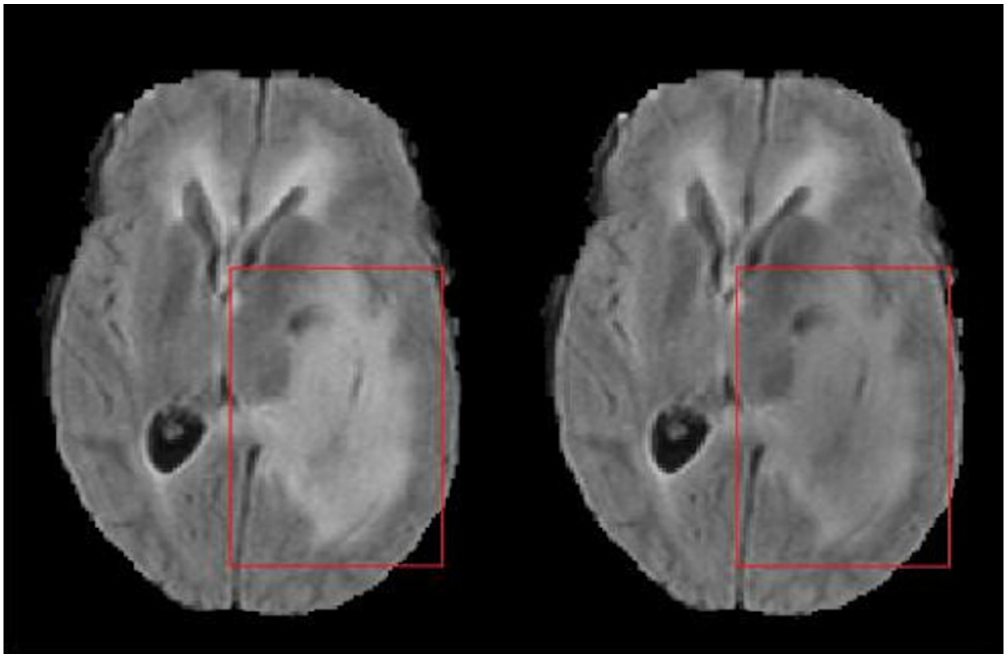
[32] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Manag. Sci.*, vol. 64, no. 3, pp. 1155–1170, Nov. 2016, doi: 10.1287/mnsc.2016.2643.

[33] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," in *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, Oct. 1993, pp. 1813–1817 vol.3. doi: 10.1109/NSSMIC.1993.373602.

Figure 1: Comparison of the field inhomogeneity correction with ANTs/Nipype (left) and SimpleITK (right). Distinct differences in the FLAIR signal intensity of tumour tissue are visible (red squares).

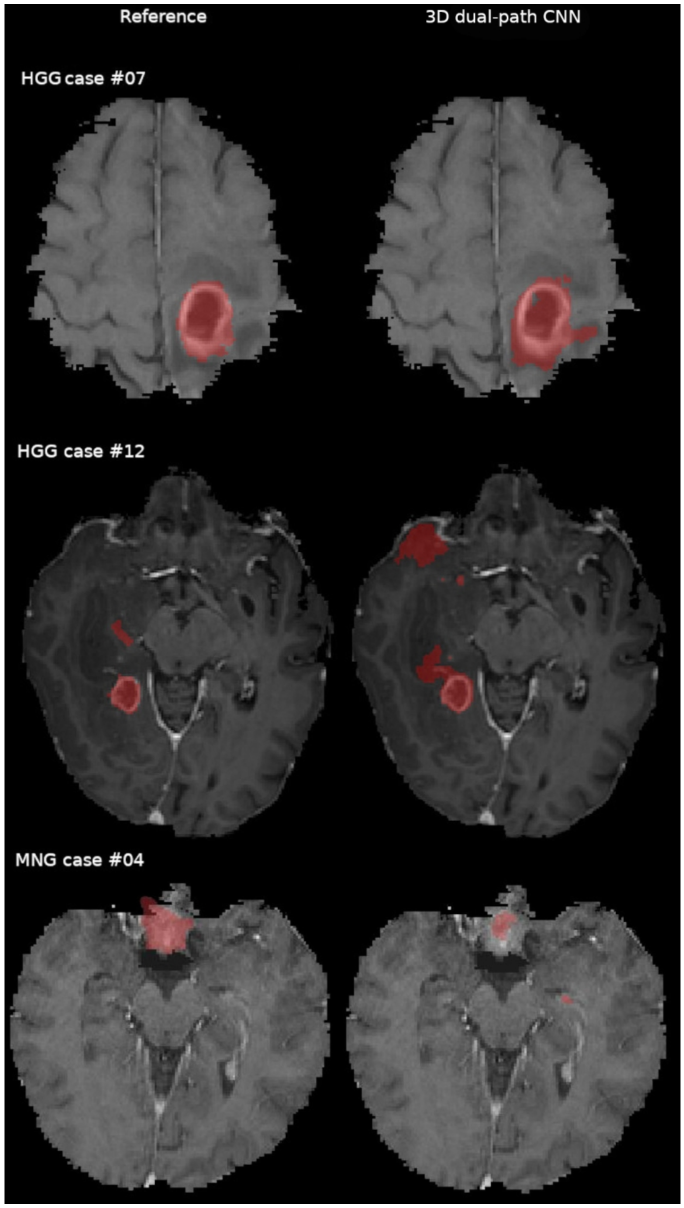
Figure 2: Comparison of the expert segmentation (reference) and the 3D dual-path CNN tumour core segmentation in the in-house data for high grade glioma (HGG) and meningioma (MNG) cases overlaid on contrast enhanced T1-weighted. Voxels misclassified by the 3D dual-path CNN are visible in HGG Cases #07 and #12 (top and middle row). The 3D dual-path CNN failed to correctly outline the tumour and included normal brain structures in the left medial temporal lobe for meningioma Case #04 (bottom row).

Figure 3: Comparison of the expert segmentation (reference) and the 3D dual-path CNN whole tumour segmentation in the in-house data for low grade glioma cases displayed overlaid on FLAIR images. Voxels misclassified by the 3D dual-path CNN are visible bilaterally in the orbit in Case #01 (top row), which should have been excluded by the skull stripping procedure. In Case #09 (middle row), the 3D dual-path CNN misclassified contralateral, sequence-depended FLAIR hyperintensities.



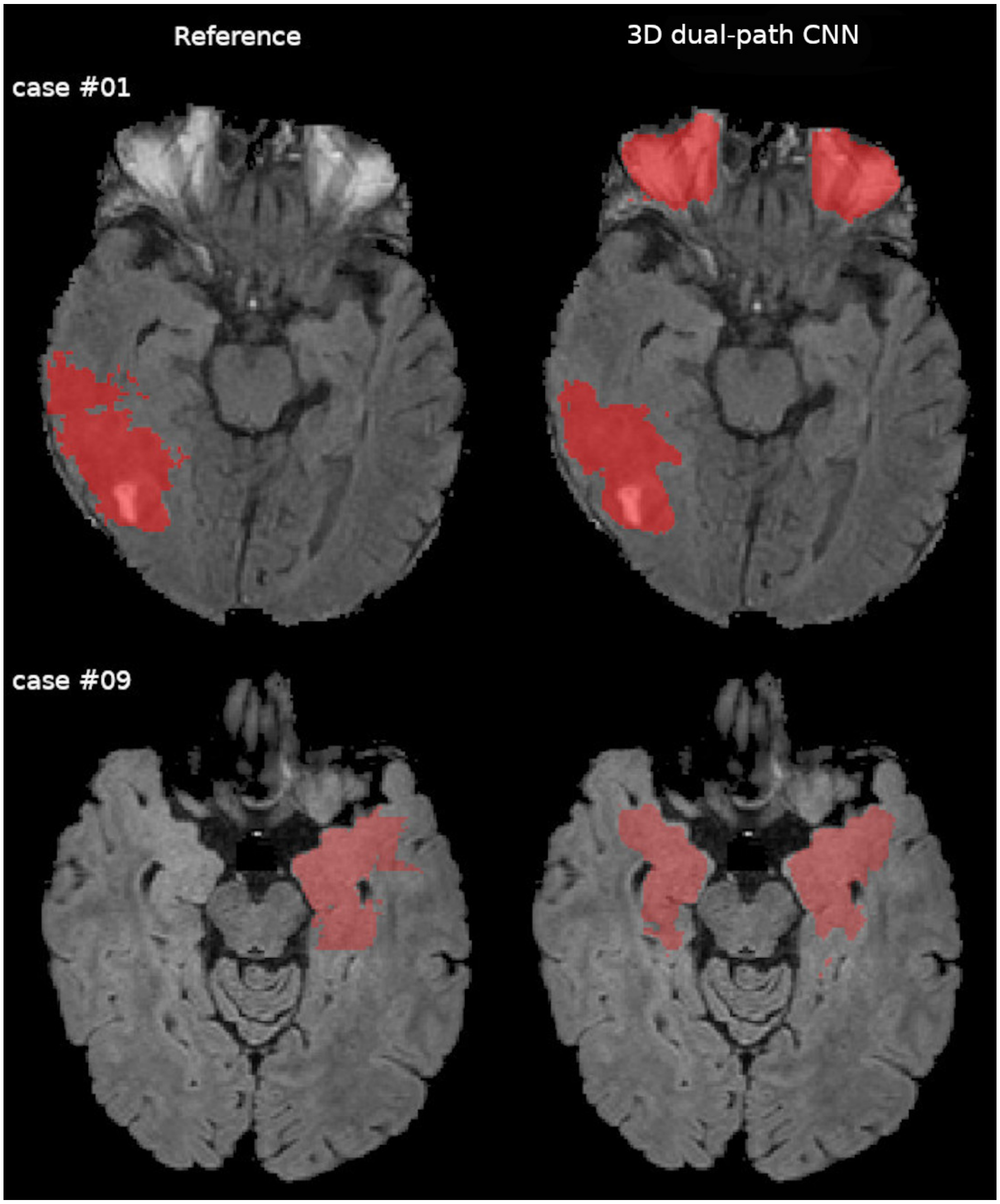
Comparison of the field inhomogeneity correction with ANTs/Nipype (left) and SimpleITK (right). Distinct differences in the FLAIR signal intensity of tumour tissue are visible (red squares).

67x44mm (300 x 300 DPI)



Comparison of the expert segmentation (reference) and the 3D dual-path CNN tumour core segmentation in the in-house data for high grade glioma (HGG) and meningioma (MNG) cases overlaid on contrast enhanced T1-weighted. Voxels misclassified by the 3D dual-path CNN are visible in HGG Cases #07 and #12 (top and middle row). The 3D dual-path CNN failed to correctly outline the tumour and included normal brain structures in the left medial temporal lobe for meningioma Case #04 (bottom row).

94x168mm (300 x 300 DPI)



Comparison of the expert segmentation (reference) and the 3D dual-path CNN whole tumour segmentation in the in-house data for low grade glioma cases displayed overlaid on FLAIR images. Voxels misclassified by the 3D dual-path CNN are visible bilaterally in the orbit in Case #01 (top row), which should have been excluded by the skull stripping procedure. In Case #09 (middle row), the 3D dual-path CNN misclassified contralateral, sequence-depended FLAIR hyperintensities.

93x112mm (300 x 300 DPI)

BMJ Open

Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059000.R2
Article Type:	Original research
Date Submitted by the Author:	16-Jun-2022
Complete List of Authors:	Gryska, Emilia; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Medical Radiation Sciences, Institute of Clinical Sciences Björkman-Burtscher, Isabella; University of Gothenburg Sahlgrenska Academy, Department of Radiology, Institute of Clinical Sciences; Sahlgrenska University Hospital, Department of Radiology Jakola, Asgeir Store; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology; Sahlgrenska University Hospital, Department of Neurosurgery Dunås, Tora; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology; University of Gothenburg Sahlgrenska Academy, Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology Schneiderman, Justin; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology Heckemann, Rolf; University of Gothenburg, MedTech West at Sahlgrenska University Hospital; University of Gothenburg Sahlgrenska Academy, Department of Medical Radiation Sciences, Institute of Clinical Sciences
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Radiology and imaging
Keywords:	Magnetic resonance imaging < RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING, Neuroradiology < RADIOLOGY & IMAGING, STATISTICS & RESEARCH METHODS

Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study

Emilia A. Gryska[1,2*], Isabella M. Björkman-Burtscher[3,4], Asgeir S. Jakola[5,6], Tora Dunås[5,7], Justin F. Schneiderman[1,5], Rolf A. Heckemann[1,2]

1 MedTech West at Sahlgrenska University Hospital, University of Gothenburg, Gothenburg, Sweden

2 Department of Medical Radiation Sciences, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

3 Department of Radiology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

4 Department of Radiology, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden

5 Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

6 Department of Neurosurgery, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden.

7 Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Sweden

* Corresponding author: Emilia A. Gryska

Postal address: MedTech West, Röda stråket 10B, Sahlgrenska University Hospital, 413 45 Gothenburg, Sweden

E-mail address: emilia.gryska@gu.se

Telephone number: +46 720304106

Keywords: Reproducibility of Results, Deep Learning, Brain Neoplasms, Magnetic Resonance Imaging

Word count: 4 094

ABSTRACT

Objectives: To determine the reproducibility and replicability of studies that develop and validate segmentation methods for brain tumours on magnetic resonance images (MRI) and that follow established reproducibility criteria; and to evaluate whether the reporting guidelines are sufficient.

Methods: Two eligible validation studies of distinct deep learning (DL) methods were identified. We implemented the methods using published information and retraced the reported validation steps. We evaluated to what extent the description of the methods enabled reproduction of the results. We further attempted to replicate reported findings on a clinical set of images acquired at our institute consisting of high and low grade glioma (HGG, LGG), and meningioma (MNG) cases.

Results: We successfully reproduced one of the two tumour segmentation methods. Insufficient description of the preprocessing pipeline and our inability to replicate the pipeline resulted in failure to reproduce the second method. The replication of the first method showed promising results in terms of Dice similarity coefficient (DSC) and sensitivity (Sen) on HGG cases (DSC=0.77, Sen=0.88) and LGG cases (DSC=0.73, Sen=0.83), however poorer performance was observed for MNG cases (DSC=0.61, Sen=0.66). Preprocessing errors were identified that contributed to low quantitative scores in some cases.

Conclusions: Established reproducibility criteria do not sufficiently emphasize description of the preprocessing pipeline. Discrepancies in preprocessing as a result of insufficient reporting are likely to influence segmentation outcomes and hinder clinical utilization. A detailed description of the whole processing chain, including preprocessing, is thus necessary to obtain stronger evidence of the generalizability of DL-based brain tumour segmentation methods and to facilitate translation of the methods into clinical practice.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- This is an independent evaluation of the reproducibility of DL-based lesion segmentation studies that follow established reporting guidelines.
- The clinical data set acquired at our institution was suitable for the replication part of the study.
- This study did not aim to enable inferences about the clinical utility of the evaluated algorithms.

INTRODUCTION

The scientific community has directed substantial efforts at developing deep-learning (DL) methods for medical image analysis. DL methods have become the default choice under the claim of superior performance to classical algorithms.[1-3] However, their outstanding performance comes at the cost of high complexity and inherent variability in model performance.[3] Consequently, assessing which model design choices determine the empirical gains is challenging.[3-5] Critics have also pointed out that scientific reporting of study designs has often been insufficient, and that the analysis of results tends to be biased towards authors' desired outcomes.[4, 6, 7] These issues present critical challenges to realizing the potential of artificial intelligence (AI) and translating promising scientific algorithms into reliable and trusted clinical decision support tools.

In our previous work,[5] we systematically explored the literature to identify whether prevalent brain lesion segmentation methods are a suitable basis for developing a tool that supports radiological brain tumour status assessment. Our findings corroborated the issues with reporting that may affect reproducibility.[5] In particular, reporting of the preprocessing steps is inadequate in many instances.

The problem has been recognized by researchers, and efforts have been made to standardize reporting practices of DL validation studies. The checklist proposed by Pineau et al.[6, 8] identifies a set of items to be reported pertaining to the presented models/algorithms, theoretical claims, data sets, code, and experimental results. The reproducibility problem in relation to the specific field of medical image segmentation was highlighted by Renard et al. in a literature review.[3] The authors present recommendations for the framework description that provides specific context for medical image segmentation. Their recommended items to be reported[3] are largely congruent with those proposed by Pineau et al.[6, 8] Renard et al.,[3] however, group their items by sources of variability in the model and evaluation framework, in contrast to grouping by scientific article section, as originally proposed by Pineau et al.[6, 8]

Furthermore, Renard et al.[3] only identified three out of twenty-nine studies included in their review to be sufficiently described according to their reproducibility recommendations. Two[9, 10] of the three were algorithms for brain tumour segmentation on magnetic resonance images (MRI). To continue our pursuit of a technically validated DL brain tumour segmentation algorithm that is suitable for clinical validation, we attempted to re-implement the two methods[9, 10].

The two DL brain tumour segmentation methods were technically validated convolutional neural networks (CNNs). Kamnitsas et al.[9] developed a 3D dual-pathway CNN with fully connected 3D conditional random fields (CRF).[11] The method will be referred to as 3D dual-path CNN in this article. The authors made the method available for independent evaluation (https://github.com/deepmedic) but did not provide a trained model. The software came with a set of configurable network parameters and requirements for the input data. The input data requirements were: images in NIfTI[12] format; images for each patient and reference labels with optional brain tissue masks (regions of interest – ROIs) had to be co-registered; all images fed to the network had to have the same voxel size; and for optimal performance, MRI signal intensities had to be standardized to have zero-mean and unit-variance within each ROI.

Pereira et al. developed a 2D single-pathway CNN, referred to as 2D single-path CNN in this article. The authors published two network architectures (HGG – high grade glioma and LGG – low grade glioma) with trained weights,[13]. The preprocessing described in the original publication consisted of bias field correction with N4ITK,[14] followed by intensity normalization[15] of each image. The input patch intensities were finally normalized with the mean and standard deviation calculated from the training patches across each sequence. A roughly similar number of patches was extracted for each class (approximately 50 000 per class for HGG to match the number of patches extracted for training as stated in the original article). The segmentation result was further processed by removing clusters of voxels smaller than a predefined threshold of 10 000 mm^3 and 3 000 mm^3 in HGG and LGG, respectively.

The aim of this study was therefore to determine the reproducibility and replicability of the two methods for brain tumour segmentation[9, 10] that Renard et al. identified as adequately reported;[3] and to evaluate whether Renard's and Pineau's reproducibility recommendations are sufficient also for the task to segment an in-house clinical data set of brain tumours.

MATERIAL AND METHODS

Overview

The study design is based on the assumption that the reproducibility items proposed by Renard et al. are sufficient for reproduction and replication. We used the definitions of reproduction and replication from the National Academies of Sciences, Engineering and Medicine,[16] which Pineau et al. also refer to.[6] Renard et al. identified two methods for brain lesion segmentation[9, 10] as adequately reported,[3] and we chose these two for the present study. Our goal was to implement the respective original methods with all processing steps and parameters and test them on the same data on which they were originally validated (reproducibility). As a measure of success, we compared quantitative results on segmentation accuracy to those reported in the original studies. We then attempted to replicate[6, 16] the findings: we performed an external validation on a clinically obtained data set from our institution.

Patient and Public Involvement

No patient involved.

Statistical analysis

We provide descriptive statistics (means, and when possible standard deviations) of segmentation evaluation metrics. The metrics we used are: Dice similarity coefficient – DSC, positive predictive value – PPV, and sensitivity.

Reproducibility analysis

Evaluated segmentation algorithms

We implemented the two previously proposed DL algorithms for brain tumour segmentation: 3D dual-path CNN[9] and 2D single-path CNN.[10] In Table 1 these algorithms are described in compliance with the reproducibility categories listed by Renard et al.,[3] together with libraries and computational parameters we used in our implementations. For our implementation, we used hyperparameters reported in the original articles. We trained the 3D dual-path CNN and tested both algorithms on a cluster with a Tesla V100 GPU (5120 cores; Nvidia Corp., Santa Clara, CA, USA), 32 GB RAM, and two 8-core Xeon Gold 6244 @ 3.60GHz processors (Intel Corp., Santa Clara, CA, USA).

Table 1: Description of the two algorithms implemented in the reproducibility analysis, 3D dual-path CNN[9] and 2D single-path CNN,[10] according to the reproducibility categories proposed by Renard et al.[3] All the parameters and versions found in the first part of the table were specified in the original articles. The selection strategy of images to respective cross-validation folds was not specified. In the part **"Our implementation middleware"**, we specify the Python version and libraries used for our implementations. CNN – convolutional neural networks, CRF – conditional random field, CV – cross-validation, DSC – Dice similarity coefficient, FC – fully connected, HGG – high grade glioma, LGG – low grade glioma.

Main category	Sub-category	3D dual-path CNN	2D single-path CNN
Algorithm/model	Description of the DL architecture	Dual-path 3D CNN with a fully connected 3D CRF.[11]	Single-path 2D CNN; two network architectures for HGG and LGG.
Dataset description	Image acquisition parameters Image size Data set size Link to the data set	BraTS 2015 dataset[17]	
Preprocessing description	Data excluded + reason	none	none
	Augmentation transformation	Sagittal reflection of images	Rotation with multiples of 90° angles
	Final sample size	Not specified	~1 800 000 for HGG ~1 340 000 for LGG
Training/validation/ testing split	Explanation if validation set not created	Training and testing sets provided by the BraTS challenge	
CV strategy + number of folds	Not specified	5-fold CV on training set (n=274)	1 subject in both HGG (n=220) and LGG (n=54)
Optimization strategy	Optimization algorithm + reference	RMSProp optimizer[18] and Nesterov's momentum[19]	Stochastic Gradient Descent and Nesterov's momentum[19]
	Hyperparameters (learning rate *a*, batch size *n*, drop-out *d*)	$a = 10^{-3}$ (halved when the convergence plateaus); $n = 10$ $d = 50\%$ (in the last 2 hidden layers)	$a_{inital} = 0.003$ $a_{final} = 0.00003$ $n = 128$ $dHGG = 0.1$ (in FC layers) $dHGG = 0.5$ (in FC layers)
	Hyperparameter selection strategy	CRF: 5-fold CV on a training subset HGG (n=44) and LGG (n=18)	Validation using 1 subject in both HGG (n=220) and LGG (n=54)
Computing infrastructure	Name, class of the architecture, and memory size	NVIDIA GTX Titan X GPU using cuDNN v5.0, 12GB	GPU NVIDIA GeForce GTX 980
Middleware	Toolbox used/in-house code + build version	Theano[20] Python 3.6.5, Tensorflow 2.0.0/1.15.0,	Theano 0.7.0[20] Lasagne 0.1dev[21] Python 2.7.10

		Nibabel 3.0.2 Numpy 1.18.2	Numpy 1.9.2
	Source code link + dependencies	https://github.com/deepmedic	http://dei-s2.dei.uminho.pt/pessoas/csilva/brats_cnn/
Evaluation	Metrics average + variations	Mean of DSC, Precision, and Sensitivity (calculated by the online evaluation system)	Boxplot and mean of DSC (calculated by the online evaluation system)
Our implementation middleware			
Python version		3.8.2	3.7.4
DL library		Tensorflow 2.2.1	Theano (git version eb6a412), Lasagne (git version 5d3c63c)
Numpy		1.18.5	1.17.3
Nibabel		3.0.2	3.2.1

Image data set used for reproducibility analysis

Both algorithms were originally validated in the 2015 Brain Tumor Segmentation Challenge (BraTS),[22] which consists of training and testing image sets of patients diagnosed with HGG and LGG. The training set contains 274 examinations (HGG n = 220, LGG n = 54). Each examination consists of T1-weighted (T1w) images before and after injection of contrast material (CM), T2w, and FLAIR (fluid-attenuated inversion recovery) images. The training data set additionally contains manual segmentations of tumour structures that serve as a criterion standard and delineate necrotic core, contrast-enhancing (CE) core, non-CE core, and oedema. For the test set containing 110 examinations the criterion standard segmentations are not publicly available. Users can upload their segmentation results to an online system[17, 23] that internally compares the results with the hidden reference to determine per-case metrics (DSC, PPV, sensitivity, and kappa). The system then returns summary measures (means and ranking position) to the user. Images in both sets are provided in .mha format and have been preprocessed with spatial normalization,[24] skull-stripping,[25] and resampling to an isotropic resolution of 1 mm^3 (linear interpolator).

Outcome parameters

We experimentally evaluated whether the two methods that Renard et al. identified as reproducible according to their proposed criteria[3] were possible to reproduce. Specifically, we examined whether enough information was given in the original articles or supplementary information for each processing step. If re-implementation did not reproduce the originally reported results, we contacted the authors directly to follow up on any missing details and added this information to the results. Pereira et al. supplied a pre-trained model;[13] for 3D dual-path CNN, we trained our re-implementation on the BraTS 2015 training data. Thereafter, we segmented the BraTS 2015 test set with both methods. We submitted the resulting segmentations to the online evaluation system[17] and recorded the summary measures returned (mean DSC, mean sensitivity, and mean PPV). Finally, we compared the summary measures with those available in the original publications.

Replication analysis

Evaluated segmentation algorithm

Only the 3D dual-path CNN was successfully re-implemented (cf. Results – Reproducibility study). External validation (replication analysis) on in-house clinical data was therefore carried out with this method. The segmentation models trained on the BraTS training data in the reproducibility analysis were applied to our dataset using a workstation with an Intel Core i7-6700HQ CPU @ 2.60 GHz processor and Nvidia GTX960M graphics card.

Image data set used for the replication analysis

The clinical in-house testing data set consisted of images from 27 cases (HGG n = 12; LGG n = 10; meningioma – MNG n = 5). The set was selected for this study from a larger sample of image data. Data were anonymized and inclusion criteria were pre-operative examinations, availability of manual expert reference segmentations, and imaging findings typical for the included types of pathology.

As in the BraTS data set, each MR examination included non-CM T1w, CM T1w, T2w, and FLAIR images. The images were provided in NIfTI[12] format. Since we used a model trained on BraTS data to segment these images, we used the BraTS-Processor module from the BraTS Toolkit[26] for preprocessing. Binary lesion segmentations had been prepared by trained personnel and revised by a senior neurosurgeon (AJ). Whole-tumour labels generated by delineation of T2/FLAIR hyperintensities were used for LGG. For HGG and MNG, the tumour core label was used, which had been delineated on CM T1w images and included CE tumour as well as any components enclosed by CE tumour. The reference segmentations were registered from the native space to the BraTS space following the transformation steps and using the registration matrices generated by the BraTS-Processor.[26]

Outcome parameters

The replicability of the 3D dual-path CNN was assessed by comparing DSC, sensitivity, and PPV derived from processing the clinical in-house data with those provided by the online system[17] during the reproducibility analysis on the BraTS test set. We visually evaluated individual cases to determine causes of segmentation errors.

Based on findings from the reproducibility and the replication analysis we reviewed recommendations on reporting items proposed by Renard et al.[3] and Pineau et al.[8] Challenges and failures in our attempts at reproduction and replication were documented and examined throughout the processes above. We then assessed and summarized these outcomes with suggested specific improvements to the reproducibility items for lesion segmentation on magnetic resonance images for brain segmentation.

RESULTS

Reproducibility study

3D dual-path CNN

BraTS data fulfilled most of the input requirements for the 3D dual-path CNN, apart from the format and the image intensity normalization. To reproduce the study, all images were converted to NIfTI format, and MR signal intensities were normalized to have zero-mean and unit-variance within each ROI. We implemented these steps using SimpleITK for image conversion and an in-house python program for signal intensity normalization. Since the BraTS images are already skull-stripped, we generated brain masks for each patient by thresholding each image to include only non-zero voxels in order to reduce the runtime of the algorithm. The only changes we made in the 3D dual-path CNN configuration file were to set the number of input channels to all four available, as described in the original article (default in the source code was CE T1w and FLAIR), and to specify not to perform validation of the available samples, as the hyperparameters had already been defined for the model. Training the algorithm took approximately 27 hours, and testing took 14.5 minutes.

The quantitative evaluation shows that our re-implementation and testing of the 3D dual-path CNN on the BraTS 2015 data set achieved comparable results to those presented in the original study (Table 2). We therefore deem the method reproducible.

Table 2: Reproducibility results on BraTS 2015 presented in the original paper for the 3D dual-path CNN[9] and for the 2D single-path CNN[10] (original) and for our independent reproducibility analysis (this work). Our analysis was carried out for high grade glioma (HGG) and low grade glioma (LGG) model parameters of the 2D single-path CNN. The results were congruent with the original analysis for the 3D dual-path CNN but they show an unsuccessful attempt to reproduce the 2D single-path CNN validation. The higher score in each column is emphasized in bold. Measures of dispersion or significance of differences were not available for the original method evaluation. CE – contrast-enhanced.

	Dice similarity coefficient			**Positive predictive value**			**Sensitivity**		
	Whole tumour	**Tumour core**	**CE tumour**	**Whole tumour**	**Tumour core**	**CE tumour**	**Whole tumour**	**Tumour core**	**CE tumour**
	3D dual-path CNN								
Original	0.85	0.67	0.63	0.85	**0.85**	**0.63**	0.88	0.61	0.66

This work	0.85	**0.68**	**0.64**	0.85	0.83	0.62	0.88	**0.64**	**0.70**
				2D single-path CNN					
Original	**0.78**	**0.65**	**0.75**	-	-	-	-	-	-
This work (HGG)	0.36	0.25	0.17	0.36	0.21	0.29	0.54	0.58	0.17
This work (LGG)	0.25	0.14	0.13	0.40	0.51	0.37	0.25	0.10	0.10

2D single-path CNN

The preprocessing description by Pereira et al. lacked certain parameters pertaining to the intensity normalization: percentile points used to create a reference histogram for each sequence and glioma grade, and intensity parameters of the training patches. Furthermore, it was not specified which model architecture was used on the BraTS 2015 test set, where the data include both HGG and LGG. Despite the missing parameters, we made an attempt to reproduce the study. We used N4ITK bias field correction (as implemented in SimpleITK) with default parameters and a histogram normalization procedure adapted from Reinhold et al.[27] We decided on this implementation instead of the corresponding function in SimpleITK, because the latter requires a reference image or histogram, neither of which was available. For the final patch-normalization step, the intensity parameters were not available, so we normalized each test image ROI to have zero-mean and unit-variance. Finally, the results were post-processed according to the procedure described by the authors. The testing time of the 2D single-path CNN was approximately 8 hours.

As the attempt was unsuccessful (results of the quantitative evaluation presented in Table 2), we approached the lead author of the method and requested the missing information. The author generously provided information on the bias field correction as well as image histogram normalization parameters.

Following this input, the N4ITK bias field correction was conducted using the implementation in ANTs[28] with the wrapper in Nipype[29] with the following parameters specified: *n_iterations* = [20, 20, 20, 10], *dimension* = 3, *bspline_fitting_distance* = 200, *shrink_factor* = 2, *convergence_threshold* = 0. A visual inspection of the field inhomogeneity correction with ANTs/Nipype and the parameters given versus SimpleITK showed signal intensity differences in the tumour region (Figure 1) that plausibly explained the failure to reproduce.

The implementation of Nyul's algorithm[15] for intensity normalization was developed in the lead author's former lab, and the author was not at liberty to share the code. Instead, the author provided percentile points and corresponding intensity landmarks for each MR sequence used in their implementation. In the original study, however, the authors trained separate sets of parameters for LGG and HGG and could not retrieve the patch intensity parameters for patch normalization. To compensate, we extracted the mean and standard deviation from the training images by collecting intensity information of patches sampled from various brain regions to ensure class balance. We imposed a condition that for a given class, a certain percentage of patch pixels are labelled as that class. The values of mean and standard deviation depended on the percentage value, and we did not succeed at finding a value that would improve the segmentation results. At this point, we decided not to pursue further efforts to reproduce the study.

Replication analysis

The replication analysis was conducted on the 3D dual-path CNN only. Quantitative results of the comparison of automatic segmented MR images collected in-house and expert delineations of the chosen tumour labels are presented in Table 3.

Table 3: 3D dual-path CNN[9] replication analysis results on in-house data for high grade glioma (HGG) cases and meningioma (MNG) cases evaluated on the tumour core and for low grade glioma (LGG) cases evaluated on the whole tumour label. DSC – Dice similarity coefficient, PPV – positive predictive value, Sen – sensitivity, Std. – standard deviation.

ID	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**Mean**	**Std.**
							HGG cases tumour core							

DSC	0.88	0.85	0.80	0.85	0.89	0.85	0.57	0.89	0.86	0.81	0.87	0.14	**0.77**	**0.22**
PPV	0.84	0.86	0.72	0.84	0.85	0.79	0.41	0.85	0.80	0.73	0.80	0.08	**0.72**	**0.23**
Sen	0.93	0.85	0.89	0.87	0.92	0.91	0.89	0.93	0.93	0.91	0.96	0.61	**0.88**	**0.09**
MNG cases tumour core														
DSC	0.84	0.80	0.56	0.09	0.77								**0.61**	**0.31**
PPV	0.89	0.72	0.41	0.60	0.66				n.a.				**0.66**	**0.18**
Sen	0.79	0.90	0.92	0.05	0.93								**0.71**	**0.38**
LGG cases whole tumour														
DSC	0.35	0.70	0.89	0.58	0.93	0.85	0.83	0.85	0.54	0.77	n.a		**0.73**	**0.18**
PPV	0.27	0.55	0.86	0.43	0.93	0.77	0.88	0.90	0.43	0.74	n.a		**0.67**	**0.24**
Sen	0.52	0.93	0.92	0.89	0.93	0.95	0.78	0.80	0.75	0.80	n.a		**0.83**	**0.13**

The average performance results of the replicability analysis using the in-house image set and the reproducibility results are compiled in Table 4 for comparison.

Table 4: Comparison of the mean results of the reproducibility (BraTS 2015 test set) and replicability (in-house image set) analysis of the 3D dual-path CNN.[9] LGG – low grade glioma, HGG – high grade glioma, MNG – meningioma, DSC – Dice similarity coefficient, PPV – positive predictive value, Sen – sensitivity.

Data set:		In-house image set		BraTS 2015 test image set
Cases:		HGG	MNG	LGG+HGG
Tumour core	DSC	0.77	0.61	0.68
	PPV	0.72	0.66	0.83
	Sen	0.88	0.71	0.64
Cases:		LGG		LGG+HGG
Whole tumour	DSC	0.73		0.85
	PPV	0.83		0.85
	Sen	0.67		0.88

The visual evaluation of individual cases revealed a variety of causes of poor performance. In HGG visual inspection of Case #07 results showed that the 3D dual-path CNN misclassified brain tissue voxels in the vicinity of the tumour core (Figure 2, top row). A similar problem was observed in Case #12 (Figure 2, middle row). The algorithm failed to segment a tumour in MNG Case #04 (Figure 2, bottom row). While the tumour location and appearance (uncharacteristic for glioma) may be the reason for a poor result, we also note that the brain mask generated in the preprocessing by BraTS Processor failed to include a part of the reference label. For LGG the algorithm achieved relatively poor results for Cases #01 and #09. The results obtained for LGG Case #01 revealed a segmentation error as a result of a preprocessing error: the brain mask included periocular tissue that was classified as tumour by the segmentation algorithm (Figure 3, top row). In LGG Case #09, the 3D dual-path CNN labelled a substantial portion of the brain that was not included in the reference segmentation (Figure 3, bottom row).

Proposed updates to the checklist

From our results we deducted that insufficient description of the preprocessing was the main obstacle to reproducing Pereira's et al.[10] results. We therefore present an updated reproducibility and replicability checklist for medical segmentation studies (Table 5).

Table 5: A suggested reproducibility and replicability checklist for automatic medical image segmentation studies. The update from the established checklists[3, 8] includes a new category **Data set preprocessing**, and a new item in Model evaluation category: **Failed cases: number and reasons.** We also regrouped the items into categories that provide a clearer structure for reporting in particular of reproducibility and replicability studies.

Data set – description of the image data set used for model development and validation:

- › Image acquisition parameters ☐
- › Data set size ☐
- › Data excluded + reason ☐
- › Link to the data set (if available) ☐

Data set preprocessing – description of the processing steps applied to the raw images before they can be fed to the segmentation model:

- › List of all processing steps and corresponding parameters developed for the implementation ☐
- › List of processing steps not included in the implementation (when segmentation model developed and validated on partially preprocessed data) ☐
- › Statement if proprietary software was used ☐
- › Link to the source code + dependencies ☐

Segmentation model – description of the model's architecture used for the segmentation:

- › Description of the model (layers, nodes, functions, etc.) ☐
- › Trained model ☐
- › Framework used to build the model + version ☐
- › Statement if proprietary software was used ☐
- › Link to the source code + dependencies ☐

Postprocessing – description of all processing steps and corresponding parameters applied to the output of the segmentation algorithm before evaluation:

- › List of all processing steps and corresponding parameters developed for the implementation ☐
- › Statement if proprietary software was used ☐
- › Link to the source code + dependencies ☐

Model development – description of the training/validation and optimization strategies:

- › Augmentation transformations and corresponding parameters used for training ☐
- › Training/validation/testing split ☐
- › Final training sample size ☐
- › CV strategy + number of folds / number of training and evaluation runs ☐
- › Optimization algorithm + reference ☐
- › Hyperparameter selection strategy ☐
- › Hyperparameters (learning rate *a*, batch size *n*, drop-out *d*) ☐
- › Link to the training source code + dependencies ☐

Computing infrastructure – description of the hardware used:

- › Name ☐
- › Class of the architecture ☐
- › Memory size ☐

Model evaluation – description of the model evaluation:

- › Metrics average + variations ☐
- › Reference segmentation source ☐
- › Failed cases: number and reasons ☐
- › Training and testing runtime ☐
- › Link to the evaluation source code or platform

DISCUSSION

Reproducibility and replicability of scientific results are the foundation of evidence-based medicine. In this work we show that current guidelines for publishing validation studies on deep-learning algorithms are incomplete. While attempting to reproduce the two studies on MR brain lesion segmentation that were identified as meeting current reproducibility recommendations,[3] we found that only one of them was reproducible based on the published information. Remarkably, even after consultation with the authors of the second method, we were not able to obtain satisfactory segmentation results with their method. Our claims of reproducibility / non-reproducibility could not be supported with advanced statistical analysis; the online evaluation system[17] (used to evaluate the segmentations in the original validation papers and our study) provides arithmetic means of the evaluation metrics without measures of dispersion. The small sample size of the in-house data along with the difference in tumour components segmented as a reference for HGG (tumour core) and LGG (whole tumour) further precludes a meaningful analysis of the statistical difference between the results obtained in the reproducibility and replicability analysis. We believe that our findings are nevertheless sufficient to support our conclusions.

We furthermore attempt to externally validate the findings reported for the 3D dual-path CNN on a set of own data. We found that the available preprocessing pipeline is not free from producing errors, which directly influences the segmentation outcome. Moreover, we observed a poorer performance of the algorithm in MNG cases. This is, however, a somewhat expected behaviour since the training set did not contain any MNG tumours. On the other hand, visual inspection also revealed potential the 3D dual-path CNN segmentation errors arising from preprocessing errors. Nonetheless, our results acquired with the BraTS-Processor and the 3D dual-path CNN are promising, and we have begun to explore the potential of this pipeline for clinical application. Unfortunately, the experience gained through this study suggests that the available algorithms are not, in their present form, ready to be implemented in clinical routines. This, despite their meeting the recommended criteria for reproducibility as outlined by Pineau et al.[6, 8] and Renard, et al.[3] Improving the reproducibility of technical validation studies of DL segmentation methods will lay a foundation for producing strong evidence for what algorithms work best, when, and why. It will furthermore facilitate creating standardized evaluation frameworks and create a solid base for implementing DL tools in clinical routines.

Reproducibility criteria

The items that Renard et al.[3] identified as necessary to reproduce a DL methodology study are divided into information about hyperparameters (optimization, learning rate, drop-out, batch size) and the data set used (training proportion, data augmentation, and validation set). All these items are indeed included in the two studies we attempted to reproduce.[9, 10] The current recommendations, however, do not sufficiently stress the importance of thorough documentation of the image preprocessing chain.

The approach to preprocessing of the training and testing data is different between the two highlighted segmentation studies. The authors of the 3D dual-path CNN guarantee optimal performance of the algorithm on images prepared for the BraTS segmentation challenge (skull stripping, spatial normalization, and resampling) with an additional intensity normalization step. The 2D single-path CNN, on the other hand, achieved its reported high accuracy after more complex preprocessing had been applied. For our study on the, intensities of the whole images were corrected for field inhomogeneity, and histograms normalized across each sequence. The final preprocessing step involved patch normalization. These procedures were not explicitly described. We requested the missing information from the authors, and while they were supportive in principle, they were unable to supply the patch intensity information. Unsurprisingly, the results show poor accuracy due to our inability to reproduce the intensity normalization procedures conducted in the original study.

The problem of insufficient reporting of the preprocessing procedures has been recognized previously.[5] While preprocessing may be less important in the context of segmentation challenges, evaluating the whole processing chain, from raw images to the final segmentation, is crucial in the context of application to independently collected data. Without the ability to reproduce the whole processing chain, meaningful method comparison and validation on external data becomes impossible.

Our findings prompt us to propose a significant modification to the previously reported reproducibility checklist by Pineau et al.[6] and Renard et al.'s guidelines.[3] We present this new checklist in Table 5. First, we add what we conclude to be a necessary and sufficient description of the preprocessing. Second, we regroup the items to provide a clearer distinction between the various elements and aspects that are involved in the algorithm development vs. the validation of the medical image segmentation tool: such a structure for providing a more transparent and easily implemented way of reporting is specifically designed to help those who seek to reproduce and replicate. More generally, these modifications are critical to improving the reproducibility and replicability of medical image segmentation methods. Since our updates are based on reproducibility and replicability of only two segmentation algorithms, we encourage researchers to comprehensively evaluate our checklist by including a broader selection of independently implemented algorithms for medical image segmentation.

Replication analysis

The external validation was conducted on locally acquired images. We cannot draw definitive conclusions regarding the 3D dual-path CNN's performance in a clinical setting as statistical analysis would not be meaningful; in the in-house data, we evaluated separately tumour core label in high grade glioma (HGG) examinations and whole tumour label in low grade glioma (LGG) examinations. The BraTS evaluations for both tumour components are, on the other hand, done on a mix of HGG and LGG cases. Because of our small sample size, we also cannot make inferences about applying deep-learning methods trained on glioma cases to other tumour cases. Our results, however, are promising. The analysis further highlighted how essential the preprocessing chain is for accurate brain tumour segmentation with the 3D dual-path CNN and likely with any other DL segmentation method.

In our pipeline, we used BraTS-Processor to take advantage of a tool that will automatically apply all the preprocessing steps that were also applied to the training set. Our analysis revealed segmentation errors that could be traced to errors in the preprocessing. Cases of errors in the skull stripping, which we observed in the in-house data, have been reported previously[30, 31] and will likely cause occasional problems in the future. Nonetheless, the processing pipeline generates segmentations that, even if erroneous in a few cases, will be easy to correct if the operator is equipped with a suitable interactive label editing tool. Developers of clinical tools should be aware of the issue and enable users to easily remove mislabelled regions.[32]

In addition to the noted preprocessing errors, we encountered another problem that likely influenced the results: the BraTS-Processor outputs images in the BraTS (MNI152[33]) space. To evaluate the automatic segmentations quantitatively, we had to transform the reference segmentations from the native space to the BraTS space as well. This resulted in visible distortions to the reference segmentations. Accordingly, the results we presented (Table 5) likely underestimate the performance of the method (BraTS-Processor + the 3D dual-path CNN) on externally acquired data. For a more accurate evaluation of a given processing pipeline, reference segmentations should be delineated on images in the BraTS space. While it may not be feasible in retrospective studies, it is a vital study design step for prospective studies.

CONCLUSIONS

Established reproducibility criteria for studies developing and validating DL lesion segmentation algorithms are not sufficient with regard to the preprocessing steps. The results of the reproducibility analysis led us to propose a new reproducibility checklist for medical image segmentation studies, especially if clinical utility of the algorithms is the goal. We further highlighted that even a fully reproducible preprocessing method is prone to errors on routine clinical images, which is likely to impair the segmentation outcome. We encourage researchers in the field of medical image segmentation to follow our modified checklist and assess it in terms of practical utility.

ETHICS APPROVAL

The data for the replication analysis were acquired under approval by the Swedish Ethical Review Authority (Dnr 702-18), which waived the requirement of informed consent.

AUTHOR CONTRIBUTIONS

EG conducted the study and led the writing of the article. RAH was the main supervisor and consultant of the study progress and design choices. JS and IB-B were co-supervising the study progress at all stages. AJ and TD provided us with the in-house collected images, reference segmentations, and design input for the external validation. All co-authors collaborated on manuscript composition and editing.

FUNDING STATEMENT

This work was supported by VGR InnovationsFonden (VGRINN-940050) and ALF funds (SU2018-03591 and ALFGBG-925851).

COMPETING INTERESTS

The authors declare that they have no competing interests.

DATA SHARING

The code generated for this study is available from https://github.com/emiliagyska/repro_study.git

ACKNOWLEDGEMENTS

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. The authors would like to thank S. Pereira for providing us with details of his study, his support, and interest in our work.

REFERENCES

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[2] J. E. Park, P. Kickingereder, and H. S. Kim, "Radiomics and Deep Learning from Research to Clinical Workflow: Neuro-Oncologic Imaging," *Korean J. Radiol.*, vol. 21, no. 10, pp. 1126–1137, Oct. 2020, doi: 10.3348/kjr.2019.0847.

[3] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, vol. 10, Aug. 2020, doi: 10.1038/s41598-020-69920-0.

[4] Z. C. Lipton and J. Steinhardt, "Research for practice: troubling trends in machine-learning scholarship," *Commun. ACM*, vol. 62, no. 6, pp. 45–53, May 2019, doi: 10.1145/3316774.

[5] E. Gryska, J. Schneiderman, I. Björkman-Burtscher, and R. A. Heckemann, "Automatic brain lesion segmentation on standard magnetic resonance images: a scoping review," *BMJ Open*, vol. 11, no. 1, p. e042660, Jan. 2021, doi: 10.1136/bmjopen-2020-042660.

[6] J. Pineau *et al.*, "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)," *ArXiv200312206 Cs Stat*, Apr. 2020, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/2003.12206

[7] B. Haibe-Kains *et al.*, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, Art. no. 7829, Oct. 2020, doi: 10.1038/s41586-020-2766-y.

[8] J. Pineau, "Machine Learning Reproducibility Checklist." https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (accessed Oct. 01, 2021).

[9] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017, doi: 10.1016/j.media.2016.10.004.

[10] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.

[11] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," *ArXiv Prepr.*, vol. arXiv:1210.5644, p. 9, 2012.

[12] Data Format Working Group, "NIfTI: — Neuroimaging Informatics Technology Initiative." https://nifti.nimh.nih.gov/ (accessed Jan. 28, 2021).

[13] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva, *Brain Tumor Segmentation using Convolutional Neural Networks in MRI Images*. [Online]. Available: http://dei-s2.dei.uminho.pt/pessoas/csilva/brats_cnn/

[14] N. J. Tustison *et al.*, "N4ITK: improved N3 bias correction," *IEEE Trans. Med. Imaging*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.

[15] L. G. Nyul, J. K. Udupa, and Xuan Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imaging*, vol. 19, no. 2, pp. 143–150, Feb. 2000, doi: 10.1109/42.836373.

[16] E. National Academies of Sciences, *Reproducibility and Replicability in Science*. 2019. doi: 10.17226/25303.

[dataset] [17] "BRATS - SICAS Medical Image Repository." https://www.smir.ch/BRATS/Start2015 (accessed Jan. 28, 2021).

[18] T. Tieleman, G. Hinton, and others, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[19] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," *ICML*, no. 3, pp. 1139–1147, 2013.

[20] F. Bastien *et al.*, "Theano: new features and speed improvements," *ArXiv12115590 Cs*, Nov. 2012, Accessed: Jun. 14, 2021. [Online]. Available: http://arxiv.org/abs/1211.5590

[21] S. Dieleman *et al.*, "Lasagne: First release." Aug. 2015. doi: 10.5281/zenodo.27878.

[22] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[23] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler, "The Virtual Skeleton Database: An Open Access Repository for Biomedical Research and Collaboration," *J. Med. Internet Res.*, vol. 15, no. 11, p. e245, 2013, doi: 10.2196/jmir.2930.

[24] L. Ibanez *et al.*, *The ITK Software Guide*. Kitware Inc (2018)., 2003.

[25] S. Bauer, T. Fejes, and M. Reyes, "A Skull-Stripping Filter for ITK," *Insight J.*, p. 859, Apr. 2012.

[26] F. Kofler *et al.*, "BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice," *Front. Neurosci.*, vol. 14, 2020, doi: 10.3389/fnins.2020.00125.

[27] J. C. Reinhold, B. E. Dewey, A. Carass, and J. L. Prince, "Evaluating the Impact of Intensity Normalization on MR Image Synthesis," *Proc. SPIE-- Int. Soc. Opt. Eng.*, vol. 10949, p. 109493H, Mar. 2019, doi: 10.1117/12.2513089.

[28] B. B. Avants, N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee, "The Insight ToolKit image registration framework," *Front. Neuroinformatics*, vol. 8, p. 44, Apr. 2014, doi: 10.3389/fninf.2014.00044.

[29] K. Gorgolewski *et al.*, "Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python," *Front. Neuroinformatics*, vol. 0, 2011, doi: 10.3389/fninf.2011.00013.

[30] F. Kellner-Weldon *et al.*, "Comparison of perioperative automated versus manual two-dimensional tumor analysis in glioblastoma patients," *Eur. J. Radiol.*, vol. 95, pp. 75–81, Oct. 2017, doi: 10.1016/j.ejrad.2017.07.028.

[31] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F. Münte, and H. Handels, "Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences," *J. Neurosci. Methods*, vol. 240, pp. 89–100, Jan. 2015, doi: 10.1016/j.jneumeth.2014.11.011.

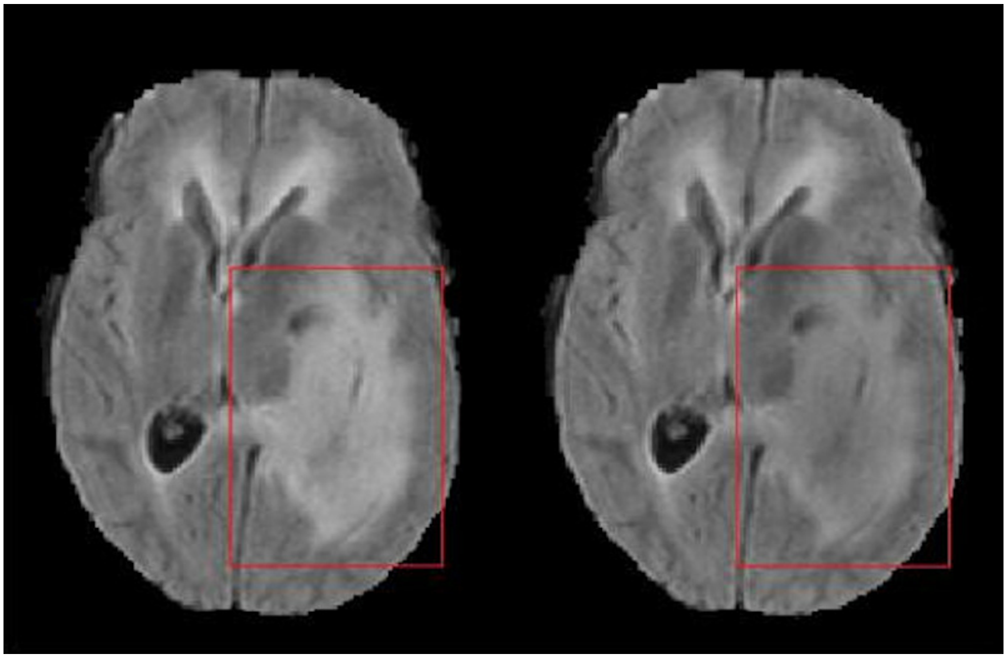
[32] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Manag. Sci.*, vol. 64, no. 3, pp. 1155–1170, Nov. 2016, doi: 10.1287/mnsc.2016.2643.

[33] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," in *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, Oct. 1993, pp. 1813–1817 vol.3. doi: 10.1109/NSSMIC.1993.373602.

Figure 1: Comparison of the field inhomogeneity correction with ANTs/Nipype (left) and SimpleITK (right). Distinct differences in the FLAIR signal intensity of tumour tissue are visible (red squares).

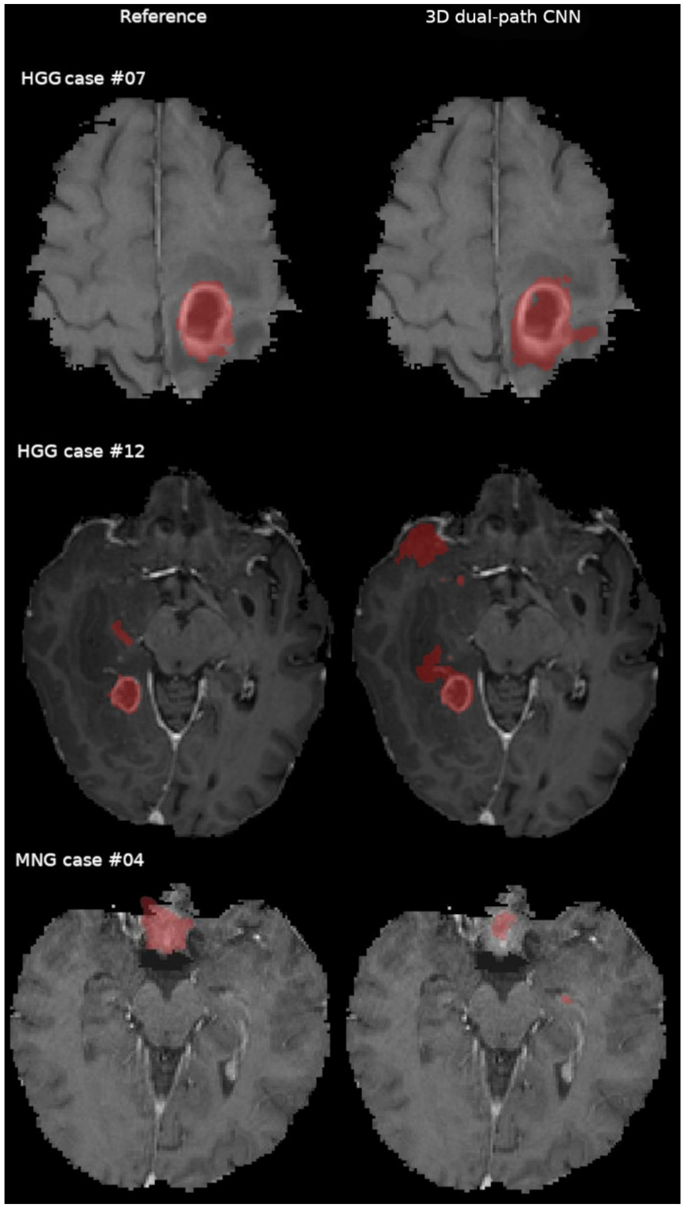
Figure 2: Comparison of the expert segmentation (reference) and the 3D dual-path CNN tumour core segmentation in the in-house data for high grade glioma (HGG) and meningioma (MNG) cases overlaid on contrast enhanced T1-weighted. Voxels misclassified by the 3D dual-path CNN are visible in HGG Cases #07 and #12 (top and middle row). The 3D dual-path CNN failed to correctly outline the tumour and included normal brain structures in the left medial temporal lobe for meningioma Case #04 (bottom row).

Figure 3: Comparison of the expert segmentation (reference) and the 3D dual-path CNN whole tumour segmentation in the in-house data for low grade glioma cases displayed overlaid on FLAIR images. Voxels misclassified by the 3D dual-path CNN are visible bilaterally in the orbit in Case #01 (top row), which should have been excluded by the skull stripping procedure. In Case #09 (middle row), the 3D dual-path CNN misclassified contralateral, sequence-depended FLAIR hyperintensities.



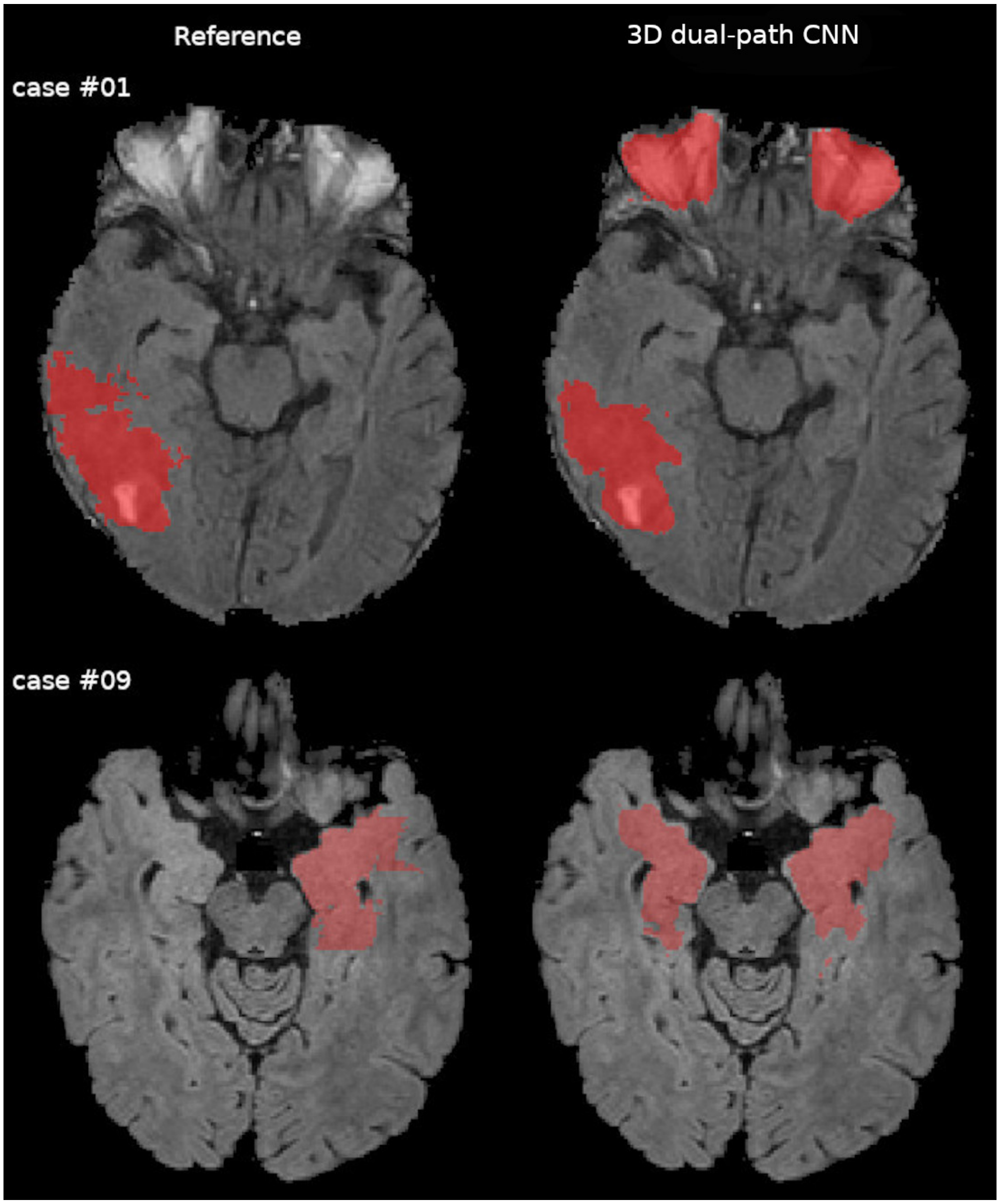
Comparison of the field inhomogeneity correction with ANTs/Nipype (left) and SimpleITK (right). Distinct differences in the FLAIR signal intensity of tumour tissue are visible (red squares).

67x44mm (300 x 300 DPI)



Comparison of the expert segmentation (reference) and the 3D dual-path CNN tumour core segmentation in the in-house data for high grade glioma (HGG) and meningioma (MNG) cases overlaid on contrast enhanced T1-weighted. Voxels misclassified by the 3D dual-path CNN are visible in HGG Cases #07 and #12 (top and middle row). The 3D dual-path CNN failed to correctly outline the tumour and included normal brain structures in the left medial temporal lobe for meningioma Case #04 (bottom row).

94x168mm (300 x 300 DPI)



Comparison of the expert segmentation (reference) and the 3D dual-path CNN whole tumour segmentation in the in-house data for low grade glioma cases displayed overlaid on FLAIR images. Voxels misclassified by the 3D dual-path CNN are visible bilaterally in the orbit in Case #01 (top row), which should have been excluded by the skull stripping procedure. In Case #09 (middle row), the 3D dual-path CNN misclassified contralateral, sequence-depended FLAIR hyperintensities.

93x112mm (300 x 300 DPI)