PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study
AUTHORS	Gryska, Emilia; Björkman-Burtscher, Isabella; Jakola, Asgeir Store; Dunås, Tora; Schneiderman, Justin; Heckemann, Rolf

VERSION 1 – REVIEW

REVIEWER	Bashir, Tariq COMSATS University Islamabad, Electrical and Computer Engineering
REVIEW RETURNED	01-Feb-2022

GENERAL COMMENTS	The novelty of the paper is not clear with respect to the proposed work. The majority of the work in the experimental result and discussion section is the comparison of the different algorithms. Statistical analysis is missing and seems no comparison with the proposed work.

REVIEWER	Gao, Bu Shijiazhuang First Hospital, Hebei Medical University, Department of Medical Research
REVIEW RETURNED	27-Feb-2022

GENERAL COMMENTS	In this study, the authors studied deep learning for automatic brain tumor segmentation on MRI. Some issues existed. 1. Abstract: Please give the full phrase for DL. The second sentence in the Methods section needs revision. 2.Introduction: In this section, please clearly describe the two segmentation methods and their advantages and disadvantages in clinical study. The background of these two methods is poorly described and we do not know what you are going to do with these two methods. More detailed description is needed 3. The method described by Pereira et al should be given in a more scientific way. For example, the DeepMedic method can be termed to Dual-path 3D CNN, and the methods described by Pereira et al can be indicated as 2D CNN. All similar terms should be changed accordingly. Do not use Pereira et al. 4. Some abbreviations used in the text should be given the full phrase at the first time of use. 5.Discussion: IN the first paragraph, the major findings should be briefly described.

REVIEWER	Lim, Gilbert National University of Singapore, School of Computing
REVIEW RETURNED	04-Mar-2022

GENERAL COMMENTS	This manuscript describes an attempt to reproduce/replicate two previously-published works on Brain MRI segmentation (following recommended reporting criteria & author correspondence if necessary), as well as an evaluation on an external (private) dataset for the successfully-replicated model, and a proposal for an expanded reporting criteria for improved replication. Reproducibility is an important consideration, especially in the medical domain. Some comments follow: 1. In the Overview subsection, it is stated that "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". The claim on "necessary" might be further justified, since it is not clear that all the items are strictly necessary (i.e. replication might have been successful omitting one/some of them) 2. In the Reproducibility analysis subsection, it is stated that "In Table 1 these algorithms are described... together with libraries and computational parameters we used in our implementations". It might be clarified as to whether these "libraries and computational parameters" were as specified by the checklists in the original works, or determined by the authors. 3. In Table 1, the "CV strategy + number of folds" item does not appear fully specified. For example, for DeepMedic, 5-fold CV on training set is listed, but this does not appear to specify the assignment to each fold. The "1 subject in both HGG and LGG" specification for Pereira et al. also appears not fully clear. 4. In Table 1, both "Hyperparameters" and "Hyperparameter section strategy" are listed. It might be clarified whether the selection strategy was actually attempted to reproduce the reported hyperparameters (and if different optimal hyperparameters were found, were they used instead), or if the reported hyperparameters were used in training the models regardless. 5. In the Pereira et al. subsection, it is stated that after the intial failure to replicate the model (with the corresponding results as reported in Table 2), "the author generously provided information on the bias field correction as well as image histogram normalization parameters", and this "plausibly explained the failure to reproduce". However, no reported attempt using this additional information appears to have been made "...we decided not to pursue further efforts to reproduce the study". Was there any reason why some attempt with the additional (partial) information was not tried (especially since the algorithm, if not the specific implementation, is known), if only to see if the performance would be closer to the reported values? 6. For the proposed updates in Table 5, the actual updated items might be highlighted, and discussed point-by-point. 7. Minor issues: (Page 5) "each imageThe" -> "each image. The" (Table 1) "int he last" -> "in the last" etc.

REVIEWER	Ghassemi, Navid Ferdowsi University of Mashhad, Computer Engineering Department

REVIEW RETURNED	13-Mar-2022

GENERAL COMMENTS	Authors have investigated one of the main issues with papers published in the field of deep learning for medical diagnosis, specifically, the reproduction of results. In my opinion, the idea is excellent, and the manuscript can contribute to the body of research in the field of study. It is good to see a manuscript that aims to solve actual problems than justifying minor improvements to previous works. However, there are a few issues with how the study is designed. In my opinion, "MATERIAL AND METHODS" and "RESULTS" need a significant change. Mainly, the reasoning behind choosing references [10] and [11] needs to be clarified. I personally think that you can see different types of issues in papers when you try to reproduce their results, and these can not be all demonstrated by reviewing two papers. This is somewhat acknowledged by the authors, too, as they have included different aspects in their final checklist, but I think to some extent, this should be demonstrated that what different (at least five distinct) papers lack, and not merely rely on one flawed work to show this. Authors have mentioned that in a prior finding, it was stated that: "three out of twenty-nine studies" had sufficient information to reproduce results. This may be accurate, but it's itself misleading. Some of the questions raised after hearing this info are: "what did those other twenty-six lack?", "Is it the same between all different papers, or each one has its own flaws?", "How much of necessary info is missing?" My suggestion is to reformat these two sections of your paper by doing these steps: - Add more than merely two references. Two can not be representative of a whole. - Select these papers to form a comprehensive representation of works done by different groups of researchers. Usually, the problems in research papers published by computer scientists are different from the ones done by physicians. I suggest selecting a few (2-3) papers from top-tier computer science conferences, such as Nips and ICML (highly cited ones) and a few other (2-3) from prestigious journals in biomedical and healthcare (such as computers in biology and medicine). - Now, after reviewing these papers and covering different aspects and views into the problem at hand, you may justify your selection of two methods for the rest of these two sections. - After the formation of the checklist, write about what was missing from the papers you've reviewed before, in a table. The rest of the results section is well-formatted and detailed in my opinion, but it might need a few changes after adding new references. Also, the checklist might change too. Lastly, I suggest adding the following item to the checklist under "Model evaluation": - Demonstration of a few samples for which the model has performed poorly.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Tariq Bashir, COMSATS University Islamabad

Comments to the Author:
R1C1: The novelty of the paper is not clear with respect to the proposed work. The majority of the work in the experimental result and discussion section is the comparison of the different algorithms.

Authors' answer to R1C1: Our reproduction and replication study might of course be seen as lacking novelty, however, we feel that the study addresses a knowledge gap we previously identified in a scoping study of the field (cited as [5] in the manuscript). In that study, we noted an abundance of published algorithms, contrasted by a scarcity of studies that independently confirm validity of existing algorithms. The latter type of work is required to achieve progress towards clinical application of advanced image analysis algorithms. We are aware that many scientific journals prefer to publish novel discoveries – or, in our field, new algorithms. The choice to submit to BMJ Open was in part based on the journal's policy to deemphasise novelty and to explicitly invite specialist studies.

R1C2: Statistical analysis is missing and seems no comparison with the proposed work.

Authors' answer to R1C2: We apologize for not explicitly describing our statistical analysis. We now have added a **Statistical analysis** subsection in the Materials and Methods where we explain that we used descriptive statistics of commonly used segmentation evaluation metrics:

> "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."

We further explain in the Discussion why no advanced statistical analysis was conducted in our study:

> "Our claims of reproducibility / non-reproducibility could not be supported with advanced statistical analysis; the online evaluation system[16] (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."

Reviewer: 2
Prof. Bu Gao, Shijiazhuang First Hospital, Hebei Medical University

Comments to the Author:
In this study, the authors studied deep learning for automatic brain tumor segmentation on MRI.

R2C1: Abstract: Please give the full phrase for DL.

Authors' answer to R2C1: We appreciate the thorough revision of our manuscript. The abbreviation is now explained.

R2C2: The second sentence in the Methods section needs revision.

Authors' answer to R2C2: We agree that the sentence was unclear. For the revised manuscript, we split it into two sentences which now read:

> "We used the definitions of reproduction and replication from the National Academies of Sciences, Engineering and Medicine[13], which were also used by Pineau et al.[6]. Renard et al. identified two methods for brain lesion segmentation [10, 11] as adequately reported [3], and we chose these two for the present study."

We hope this addresses the reviewer's criticism.

R2C3: Introduction: In this section, please clearly describe the two segmentation methods and their advantages and disadvantages in clinical study. The background of these two methods is poorly

described and we do not know what you are going to do with these two methods. More detailed description is needed

Authors' answer to R2C3: We revised the introduction section according to the reviewer's suggestion. We provide a description of DeepMedic and Pereira et al.'s method (referred to as the 3D dual-path CNN and the 2D single-path CNN respectively, according to the suggestion from R2C4), and clarify how and why we used them in our study. We further highlight that the papers where the methods were originally proposed provide validation at a technical level, whereas our study evaluates eligibility of the two methods for clinical studies.

R2C4: The method described by Pereira et al should be given in a more scientific way. For example, the DeepMedic method can be termed to Dual-path 3D CNN, and the methods described by Pereira et al can be indicated as 2D CNN. All similar terms should be changed accordingly. Do not use Pereira et al.

Authors' answer to R2C4: We agree with the reviewer and made changes accordingly, now referring to the algorithms as "3D dual-path CNN" and "2D single-path CNN".

R2C5: Some abbreviations used in the text should be given the full phrase at the first time of use.

Authors' answer to R2C5: Thank you for pointing out the omissions. We revised the manuscript according to the comment.

R2C6: Discussion: IN the first paragraph, the major findings should be briefly described.

Authors' answer to R2C6: Thank you for the suggestion. We agree that major findings should be summarized early on in the Discussion and we have revised the first paragraph accordingly. It now reads:

> "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."

Reviewer: 3
Dr. Gilbert Lim, National University of Singapore

Comments to the Author:
This manuscript describes an attempt to reproduce/replicate two previously-published works on Brain MRI segmentation (following recommended reporting criteria & author correspondence if necessary), as well as an evaluation on an external (private) dataset for the successfully-replicated model, and a proposal for an expanded reporting criteria for improved replication. Reproducibility is an important consideration, especially in the medical domain.

Authors' answer: Thank you for this summary and favourable assessment.

Some comments follow:

R3C1: In the Overview subsection, it is stated that "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". The claim on "necessary" might be further justified, since it is not clear that all the items are strictly necessary (i.e. replication might have been successful omitting one/some of them).

Authors' answer to R3C1: Agreed: we only evaluate sufficiency of these items, not whether they are necessary. We revised the sentence accordingly and it now reads:

"The study design is based on the assumption that the reproducibility items proposed by Renard et al. are sufficient for reproduction and replication."

R3C2: In the Reproducibility analysis subsection, it is stated that "In Table 1 these algorithms are described... together with libraries and computational parameters we used in our implementations". It might be clarified as to whether these "libraries and computational parameters" were as specified by the checklists in the original works, or determined by the authors.

Authors' answer to R3C2: We agree and have now clarified the description of Table 1 to indicate which parameters were specified by the authors of the original works and which were used in our implementation:

"All the parameters and versions found in the first part of the table were specified in the original works. In the part "Our implementation middleware**"**, we specify the Python version and libraries used for our implementations."

We also added the libraries and versions specified in the original articles in Table 1. The relevant part of the table with the changes highlighted is pasted below:

Middleware	Toolbox used/in-house code + build version	Theano[19] Python 3.6.5, Tensorflow 2.0.0/1.15.0, Nibabel 3.0.2 Numpy 1.18.2	Theano 0.7.0[19] Lasagne 0.1dev[20] 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/

R3C3: In Table 1, the "CV strategy + number of folds" item does not appear fully specified. For example, for DeepMedic, 5-fold CV on training set is listed, but this does not appear to specify the assignment to each fold. The "1 subject in both HGG and LGG" specification for Pereira et al. also appears not fully clear.

Authors' answer to R3C3: Thank you for noting this omission. We have now added the size of the training set in Table 1 so that these parameters are meaningful:

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[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); $n = 10$ $d = 50\%$ (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)

R3C4: In Table 1, both "Hyperparameters" and "Hyperparameter section strategy" are listed. It might be clarified whether the selection strategy was actually attempted to reproduce the reported

hyperparameters (and if different optimal hyperparameters were found, were they used instead), or if the reported hyperparameters were used in training the models regardless.

Authors' answer to R3C4: We agree with the suggestion. We have added a clarification to the Materials and Methods that reads:

> "For our implementation, we used hyperparameters reported in the original articles."

R3C5: In the Pereira et al. subsection, it is stated that after the initial failure to replicate the model (with the corresponding results as reported in Table 2), "the author generously provided information on the bias field correction as well as image histogram normalization parameters", and this "plausibly explained the failure to reproduce". However, no reported attempt using this additional information appears to have been made "...we decided not to pursue further efforts to reproduce the study". Was there any reason why some attempt with the additional (partial) information was not tried (especially since the algorithm, if not the specific implementation, is known), if only to see if the performance would be closer to the reported values?

Authors' answer to R3C5: We omitted the description of these further attempts from the Results by mistake and only mentioned them in the Discussion. We are grateful to the Reviewer for pointing out this flaw. We have updated the Results section with the following:

> "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 and did not result in satisfactory results."

R3C6: For the proposed updates in Table 5, the actual updated items might be highlighted, and discussed point-by-point.

Authors' answer to R3C6: We followed the Reviewer's suggestion to indicate what constituted the updates in Table 5. We added to the caption of Table 5:

> "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."

We also reiterated and included more detailed reasoning for the regrouping of the items in the Discussion:

> "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."

R3C7: Minor issues:
(Page 5) "each imageThe" -> "each image. The"
(Table 1) "int he last" -> "in the last"
etc.

Authors' answer to R3C7: We apologize for these errors. We have thoroughly re-reviewed the manuscript for typographical errors. The ones pointed out have, of course, been changed according to the suggestions.

Reviewer: 4
Dr. Navid Ghassemi, Ferdowsi University of Mashhad

Comments to the Author:

R4C1: Authors have investigated one of the main issues with papers published in the field of deep learning for medical diagnosis, specifically, the reproduction of results. In my opinion, the idea is excellent, and the manuscript can contribute to the body of research in the field of study. It is good to see a manuscript that aims to solve actual problems than justifying minor improvements to previous works.

Authors' answer R4C1: We greatly appreciate the encouraging comment.

R4C2.1: However, there are a few issues with how the study is designed. In my opinion, "MATERIAL AND METHODS" and "RESULTS" need a significant change. Mainly, the reasoning behind choosing references [10] and [11] needs to be clarified.

Authors' answer to R4C2.1: We apologize that we failed to convey what the scope and aim of our study were. Our mistake logically led to a misunderstanding of what we intended to do followed by a reasonable suggestion to improve the methodology of the study. To correct our mistake, we have now added a paragraph in the introduction that explains the choice of the two methods for this study (edited parts highlighted in yellow):

> "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."

(…)

> "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[10, 11] 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[10, 11]."

R4C2.2: I personally think that you can see different types of issues in papers when you try to reproduce their results, and these cannot be all demonstrated by reviewing two papers. This is somewhat acknowledged by the authors, too, as they have included different aspects in their final checklist, but I think to some extent, this should be demonstrated that what different (at least five distinct) papers lack, and not merely rely on one flawed work to show this. Authors have mentioned that in a prior finding, it was stated that: "three out of twenty-nine studies" had sufficient information to reproduce results. This may be accurate, but it's itself misleading. Some of the questions raised after hearing this info are: "what did those other twenty-six lack?", "Is it the same between all different papers, or each one has its own flaws?", "How much of necessary info is missing?"

My suggestion is to reformat these two sections of your paper by doing these steps:

- Add more than merely two references. Two cannot be representative of a whole.

- Select these papers to form a comprehensive representation of works done by different groups of researchers. Usually, the problems in research papers published by computer scientists are different from the ones done by physicians. I suggest selecting a few (2-3) papers from top-tier computer science conferences, such as Nips and ICML (highly cited ones) and a few other (2-3) from prestigious journals in biomedical and healthcare (such as computers in biology and medicine).

- Now, after reviewing these papers and covering different aspects and views into the problem at hand, you may justify your selection of two methods for the rest of these two sections.

- After the formation of the checklist, write about what was missing from the papers you've reviewed before, in a table. The rest of the results section is well-formatted and detailed in my opinion, but it might need a few changes after adding new references. Also, the checklist might change too.

Authors' answer to R4C2.2: The reviewer is correct to propose these steps for a study that aims to comprehensively evaluate existing reproducibility checklists. Since this was not the aim of our work, we did not implement these steps in the current study. We recognize nevertheless the value of conducting a comprehensive evaluation as proposed by the reviewer. We added this suggestion in the Discussion as a potential direction for future studies:

> "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."

R4C3: Lastly, I suggest adding the following item to the checklist under "Model evaluation":
- Demonstration of a few samples for which the model has performed poorly.

Authors' answer to R4C3: Good point: the "failed cases" are often not mentioned but are potentially a valuable source of information. We added the information in Table 5 under the category **Model evaluation:**

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

VERSION 2 – REVIEW

REVIEWER	Lim, Gilbert National University of Singapore, School of Computing
REVIEW RETURNED	13-May-2022

GENERAL COMMENTS	We thank the authors for addressing almost all our previous concerns. The minor remaining concern rests on the cross-validation set used. Were the exact training splits (i.e. which image belongs to which training/test set in each cross-validation fold) known, or was this independently randomized? This might be briefly clarified.

REVIEWER	Ghassemi, Navid Ferdowsi University of Mashhad, Computer Engineering Department
REVIEW RETURNED	25-May-2022

GENERAL COMMENTS	The authors have answered all of my comments, and I recommend it for publication in its current form.

VERSION 2 – AUTHOR RESPONSE

Reviewer 3, comment 1: We thank the authors for addressing almost all our previous concerns. The minor remaining concern rests on the cross-validation set used. Were the exact training splits (i.e. which image belongs to which training/test set in each cross-validation fold) known, or was this independently randomized? This might be briefly clarified.

Our response: We addressed the concern in the description of table 1 that reads now:

> "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."

Reviewer 4, comment 1: The authors have answered all of my comments, and I recommend it for publication in its current form.

Our response: We thank the reviewer for this recommendation.

VERSION 3 – REVIEW

REVIEWER	Lim, Gilbert National University of Singapore, School of Computing
REVIEW RETURNED	24-Jun-2022

GENERAL COMMENTS	We thank the authors for addressing our remaining concern.