2D and 3D U-Nets for skull stripping in a large and heterogeneous set of head MRI using `fastai`

Satheshkumar Kaliyugarasan[1,2,*], Marek Kociński[1,3,4,*], Arvid Lundervold[1,3,*], Alexander Selvikvåg Lundervold[1,2,*], for the Alzheimer's Disease Neuroimaging Initiative[**], and for the Australian Imaging Biomarkers and Lifestyle flagship study of ageing[***]

[1]Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland University Hospital, Bergen, Norway
[2]Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway
[3]Dept. of Biomedicine, University of Bergen, Norway
[4]Institute of Electronics, Lodz University of Technology, Poland
[*]All authors contributed equally to the work

[**]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
`http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

[***]Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database. The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at `www.aibl.csiro.au`.

Abstract

Skull stripping in brain imaging is the removal of the parts of images corresponding to non-brain tissue. Fast and accurate skull stripping is a crucial step for numerous medical brain imaging applications, e.g. registration, segmentation and feature extraction, as it eases subsequent image processing steps. In this work, we propose and compare two novel skull stripping methods based on 2D and 3D convolutional neural networks trained on a large, heterogeneous collection of 2777 clinical 3D T1-weighted MRI images from 1681 healthy subjects. We investigated the performance of the models by testing them on 927 images from 324 subjects set aside from our collection of data, in addition to images from an independent, large brain imaging study: the IXI dataset ($n = 556$). Our models achieved mean Dice scores higher than 0.978 and Jaccard indices higher than 0.957 on all tests sets, making predictions on new unseen brain MR images in approximately 1.4s for the 3D model and 12.4s for the 2D model. A preliminary exploration of the models' robustness to variation in the input data showed favourable results when compared to a traditional, well-established skull stripping method. With further research aimed at increasing the models' robustness, such accurate and fast skull stripping methods can potentially form a useful component of brain MRI analysis pipelines.

1 Introduction

Magnetic resonance imaging of the brain

Magnetic resonance imaging (MRI) is a medical imaging technology (modality) used in radiology to acquire information in space and time about structure (anatomy) and

This paper was presented at the NIK-2020 conference; see `http://www.nik.no`.

function (physiology) of tissues and organs in the body. MRI scanners use a combination of strong magnetic fields, magnetic field gradients for spatial encoding and decoding of nuclear spin populations, typically protons (e.g. water) in different chemical and microstructural environments, radio waves, and image reconstruction algorithms working in complex-valued Fourier space. This is used to generate 2D, 3D, 3D+time, or even higher dimensional images of organs, providing information about tissue states and physiological and biochemical processes. Among the most frequent organs subject to MRI examinations is the brain. There are several reasons for this: (i) MRI measurements can collect unsurpassed rich and detailed soft tissue information from the living brain in health and disease with little risk for the patient, and at multiple times during a disease process; (ii) compared to most other parts of the body the brain is an organ for which invasive biopsies (tissue samples) are rarely indicated, for obvious reasons; (iii) the brain within the skull can be kept rather stationary in the head coil during MR measurement time (total examination time is usually 15 - 45 min) in contrast to e.g. the beating heart or abdominal organs that move due to respiration and pulsations causing displacements and movement artifacts that are challenging to correct for, and finally (iv) most of the new MRI measurement techniques (e.g. high resolution structural MRI, diffusion MRI and functional MRI) and advanced image analysis developments tend to first enter the brain and neuro-imaging field before being adapted and applied to other organs.

Deep learning in brain imaging

Recent years' surge of interest in image analysis approaches based on *deep learning* is a case in point [1]. Considerable advances in computers' ability to extract meaningful, actionable information from complicated and heterogeneous datasets have resulted in remarkable achievements in general computer vision, natural language processing, data synthesis, sequence analysis, robotics, the analysis of tabular structured datasets, and more. Driven by these advances, the field of artificial intelligence is experiencing a tremendous amount of attention from researchers, industry, funding agencies, government and entrepreneurs, leading to rapid progress in methods, applications and products. Artificial intelligence in medicine has a long history, dating back to at least the early 1970s[1], but the field hasn't yet had a broad impact on medical practice [3]. Recently, the possibilities of using deep learning on medical data has proven to be highly potent, leading to a torrent of publications across many medical disciplines: radiology, psychiatry, dermatology, pathology, ophthalmology, cardiology, electronic health records, drug discovery, genome sequencing, and much more. See the continuously updated review `https://greenelab.github.io/deep-review`.[2]

What is skull stripping?

Skull stripping, also called brain extraction, is the task of extracting the cerebrum and the cerebellum, including cerebrospinal fluid (CSF) in the subarachnoid space from a given 3D MRI head acquisition (cf. Fig. 1). The brainstem is cut according to a specified level (e.g. distal part of medulla oblongata), assuming this level of the central nervous system is located within the field of view of the image. See the white arrow in Fig. 2 d) for an illustration. Inclusion of extra-dural tissue, e.g. skull, scalp, muscle or fat, or exclusion of brain parenchyma proper, e.g. cuts into gray matter or white matter, are considered skull stripping failures.

[1] e.g. the Mycin system of [2] aimed at identifying bacterial infections and recommending antibiotics

[2] A soon-to-be-updated published version of the survey from 2018 is available in [4]

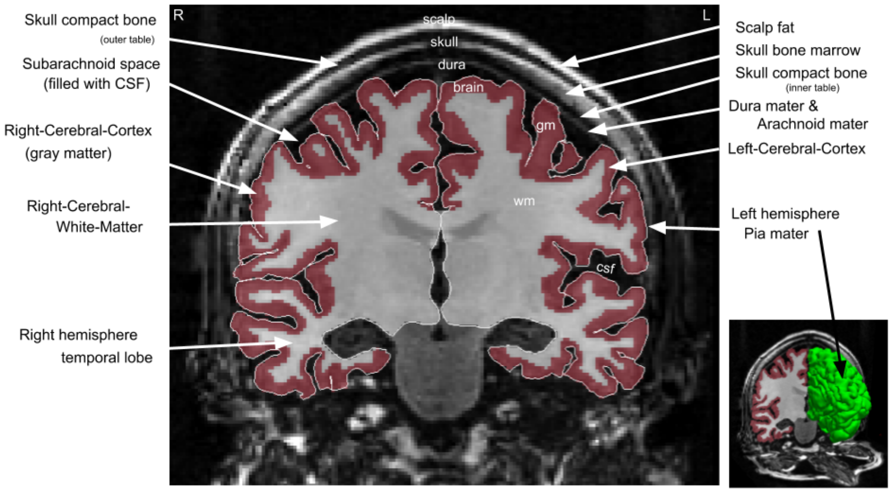

Figure 1: *Anatomy of the head related to the brain extraction task.* A coronal slice from a 3D T1-weighted (T1w) MRI recording from the head showing the different anatomical structures relevant to the segmentation task of skull stripping or brain extraction (data from [5]). Fully automated segmentation of brain (ribbon) including gray matter (gm) and white matter (wm) of the left and right hemisphere and the outer pial boundary of the brain (white continuous tracing and the surface rendering in the small insert) was performed using `Freesurfer v.7.1.1`. CFS = cerebrospinal fluid. A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure1`.

Skull stripping is important

Skull stripping is essentially *a region of interest (ROI) segmentation procedure* for subsequent analysis of structural and functional image-derived properties, spatially restricted to the brain, the brainstem (midbrain, pons, medulla oblongata) and the cerebellum. Considering signal intensities, several tissues outside the skull will have intensity distributions that overlap with principal tissue types within the brain. E.g. skeletal muscle in the head have very similar signal intensities in T1w MRI acquisitions to those observed in cerebral gray matter, and blood perfusion time courses or water diffusion properties outside the skull might have similar shape or characteristics as observed within the brain. Thus, for visualization purposes and for quantification (e.g. mean value of an imaging-derived parameter with in the brain) a skull stripping procedure is essential. Moreover, a spatially meaningful restriction of a 2D, 3D or 4D (multispectral 3D or 3D+time) image will help subsequent segmentation algorithms in further spatial refinement and increased anatomical and functional granularity within the brain (e.g. tissue classification in health and disease, or functional connectivity analysis from fMRI recordings assuming all nodes in a network graph are located within the brain, or a sub-region of the brain).

Skull stripping is difficult

There are many sources of difficulty for brain MRI image analysis methods, ranging from scanner and acquisition protocol variation, to subject motion and varying head position in the coil. One important challenge is the presence of a *bias field*. This is is usually perceived as a low-frequency, smooth variation of intensities across a slice image that degrades the MRI recording. The same tissue occurring at different locations within the image can have different signal intensity, invalidating the piecewise constant property of ideal images. Such MRI bias field is caused by an improper image acquisition process, such as radio frequency coil (B_1) non-uniformity or inhomogeneity of the main magnetic field (B_0), this being more prevalent in older MRI scanners or in ultra high-field ($B_0 \geq 7$ T) scanners. Trained radiologist are hardly influenced by this, as they easily compensate for this non-biological intensity variation in image regions. However, the bias field can pose a difficulty for quantitative image analysis algorithms assuming a spatially invariant relation between signal intensity (gray level) distribution and underlying tissue type or state. In the context of skull stripping and brain segmentation, a bias field correcting algorithm is therefore typically applied as a preprocessing step. See Fig. 2 for an example.

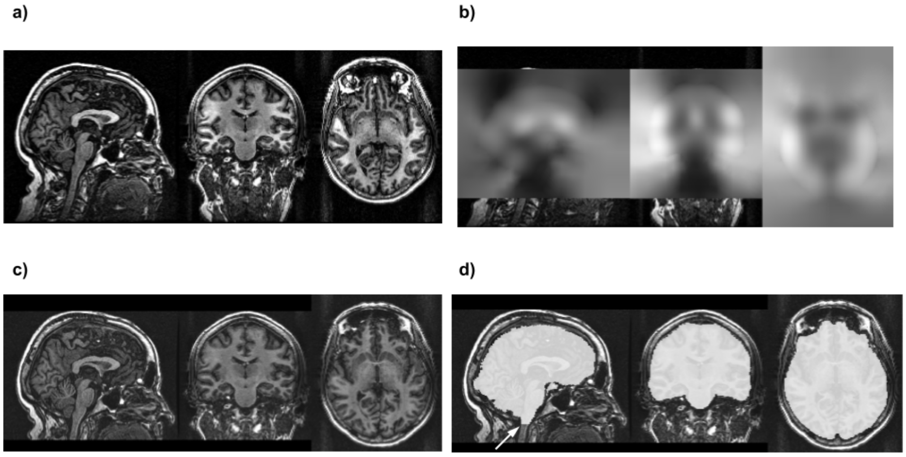

Figure 2: Bias field correction and skull stripping using `fsl_anat` (same subject as in Fig. 1). a) Original acquisition with substantial signal intensity inhomogeneity due to the presence of a bias field. b) The bias field estimate (low-frequency and smooth intensity variation across the image). c) Bias field corrected (i.e. suppressed) image. d) Skull stripping of the bias field corrected image. Arrow indicate the cut of the brainstem, defining the lower boundary of the brain.

Related work

As it is such a fundamental task in brain image analysis there has been a lot of research into skull stripping since the advent of brain MRI image analysis, leading to many proposed methods. These can be roughly categorized into machine learning- and conventional non-machine learning-based approaches. Among conventional methods there's a wide variety of approaches, based on surfaces, morphology, image intensity, templates, or hybrids of these, resulting in a number of well-established, frequently used skull stripping tools in brain image analysis pipelines [6], e.g. the Brain Extraction Tool (BET) [7] of the FMRIB Software Library (FSL), v.6.0 [8], `antsBrainExtraction` from Advanced Normalization Tools (ANTs) [9] and the `3dSkullStrip` tool in AFNI [10]. The machine learning approaches are either based on “classical" machine learning models, e.g. SVMs, region growing, active contours, or based on deep neural networks. It is the latter category that our own work belongs. Two recent illustrative examples of related approaches are presented below.

In [11] the authors developed an automated skull stripping algorithm called **HD-BET** that works for pre-contrast T1w, post-contrast T1w, T2w and FLAIR sequences. Their three-dimensional U-Net-like CNN was trained on 6.586 MR images from 1568 exams of 372 patients collected at 25 different institutions in the EORTC-2610 study. As ground truth brain masks they used BET as a starting point then had a radiologist do visual inspection and corrections (i.e. a single rater). During training, the images were resampled to isotropic spacing of 1.5mm^3 and patches of size 128^3 voxels were randomly sampled from the four different input modalities before being fed to the model. They used a relatively large set of data augmentation techniques: randomly mirroring the image patches along all axes, scaling, rotation and elastic deformations, gamma augmentation, adding additive Gaussian noise, and Gaussian blurring. They scored their model on five independent test sets: one created using the data from 12 institutions in the EORTC-2610 study not present in their training data, and the three openly available datasets LPBA40 from LONI, NFBS and CC-359, for which manually constructed ground truth masks are available. On T1w images from the EORTC-2610 study, their model had a median Dice score of 97.6 (97.0-98.0 IQR) and a median Hausdorff distance of 3.3 (2.2-3.3 IQR). On the three openly available datasets their model obtained a Dice score of 97.5 (17.4-97.7), 98.2 (98.0-98.4), 96.9 (96.7-97.1), respectively, when compared to the provided ground truth reference masks.

The **CompNets** of [12] are multi-pathway two-dimensional U-Net-like models with an embedded W-Net-like component [13], tasked with extracting information from both the brain and non-brain tissue in the input images. Their models were trained on T1w images from 406 subjects aged 18-96 from the OASIS dataset, using the brain masks provided with the OASIS dataset release as ground truth labels. All images were of size 256^3, and their models were trained using 2D slices of the 3D images, with no data augmentation. After making predictions, the masks for each slice were stacked into 3D images. No postprocessing of the resulting predicted brain masked was performed. In a two-fold cross-validation setup were the OASIS subjects are equally divided into two chunks for training and testing, their best model achieved an average Dice score of 98.27 ± 0.30.

Main contributions of our work

1. We construct high-performing skull-strip models from a large, heterogeneous dataset sourced from seven different brain imaging studies. Our results compare favourably with other state-of-the-art models based on deep learning, although direct comparisons between methods are difficult because of the lack of an agreed upon ground truth. This is an issue we discuss in our work.

2. With a novel combination of the `MONAI` deep learning library and our own extension of the `fastai` library to 3D problems, we are able to use multiple interesting state-of-the-art techniques for the construction and training of models.

3. We evaluate the performance of our models on data completely unseen during model construction. Some of which were gathered by a brain imaging study using different combinations of scanners and scanning protocols than those represented in the training data.

4. Once a skull stripping approach reaches a certain average performance level, then arguably the robustness to variation in the input data becomes more important than increased average performance. Our work indicates that CNN-based approaches to skull stripping have some robustness advantages over traditional methods.

5. The data used in our study is available to researchers through various project websites (linked below), easing reproductions and comparisons with other skull stripping methods.

2 Methods and materials

Image datasets

We compiled a large collection of T1w images of healthy volunteers from a number of different data sources[3]: ADNI, AIBL, IXI, PPMI, SLIM, Calgary-Campinas and SALD. This is a highly heterogeneous collection, involving a large number of subjects, scanners and scanner protocols, image sizes and voxel spacings, making it a challenge for any model to make predictions, but also leading to models that are more robustness to such variation. The studies were approved by the relevant Institutional Review Boards at each site and informed consent was obtained from all subjects prior to enrollment. All methods were carried out in accordance with relevant guidelines and regulation. Part of the data material used was sourced from the ADNI database. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of Mild Cognitive Impairment (MCI) and early Alzheimer's Disease (AD) [14]. We also used data collected by the AIBL study group. AIBL study methodology has been reported previously [15].

Note that we selected the subjects that were marked as healthy throughout all these longitudinal studies.

Image preprocessing and label generation

The steps used to automatically construct ground truth labels were performed using a set of well-established, validated tools. There are no manual steps in this process, making it easy to scale our approach to a large number of images. The DICOM recordings were first converted to NIfTI data format using `dicom2niix` (v1.0.20190902, [16]). To reduce the effect of scanner variation, we performed bias field correction, before producing masks indicating the location of the brain. These last two steps were done using a combination of multiple tools from the FMRIB Software Library (FSL) v.6.0 [8], collected in the `fsl_anat` pipeline[4]: (i) reorientation to match the MNI152 standard template orientation using `reorient2std` in FSL, (ii) bias field correction using `FAST` [17], (iii) linear and nonlinear registration to standard MNI152 space using `FLIRT` and `FNIRT` [18, 19], from which the brain was extracted [7]. The entire set of preprocessing steps takes on average less than 10 minutes per volume on a standard workstation computer (e.g. on an Intel Core i7-7700K CPU running Ubuntu 18.04 GNU/Linux). Finally, all volumes were resampled to isotropic $1.0 \times 1.0 \times 1.0$ mm^3 voxel size with the use of the `Convert3D Tool`.

The preprocessed images and ground truth masks were used to create training and testing datasets. Our 2D and 3D setups were based on exactly the same underlying subjects and images, placed in common training and test sets. The training dataset for our 3D model contained 2791 NIfTI files, while the two test sets, tes and IXI, consisted of 934 and 561 images, respectively. For the 2D approach, each 3D volume ($\sim$ 170 axial slices) was split into a set of 2D axial cross-sections. The total number of image files used to train the 2D model was then 469.116, while the two test datasets contained 157.036 and 95.520 image files, respectively.

[3] Links to all the data sources used in this work can be found here: `https://github.com/MMIV-ML/Skull-stripping-NIK2020`

[4] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat`

	2D U-Net	**3D U-Net**
Optimizer	Adam	Adam
Base learning rate	0.0001	0.01
Loss function	Binary Cross-Entropy	Based on the Jaccard Index
Image size	128 x 128	160 x 160 x 92
Data augmentation	Random rotation [−15°, 15°] and scaling [1.0, 1.05]	Random rotation [−10°, 10°] and scaling [1.0, 1.1]
Dropout	0.5	0.5
Weight decay	0.01	0.01
Batch size	128	8
GPU	NVIDIA Titan RTX 24GB	4 x NVIDIA Tesla V 1000 32GB

Table 1: Experimental settings for our 2D and 3D U-Net models.

Constructing and training the 2D and 3D models

We used two different U-Net models in this work: (i) A dynamic 2D U-Net using a ResNet-34 model pre-trained on the ImageNet dataset for image feature extraction (encoder) and PixelShuffle [20] with ICNR initalization [21] for upsampling (decoder), implemented in the PyTorch-based `fastai v1`; (ii) a 3D U-Net implemented using MONAI, a PyTorch based library for deep learning in healthcare imaging, and trained using our own extension of the `fastai` library. The computer vision implementations of the `fastai` library are mostly tailored to 2D imaging. We adapted the library to 3D MR images by constructing new data loaders and data augmentation capabilities, as well as adapting various 2D-specific functionality in the `fastai` library. This enables the use of custom 3D CNNs while still supporting the highly impactful training techniques of `fastai`. This includes the learning rate finder to find the optimum learning rate and the one-cycle policy (e.g., learning rate changes during the training, related to what is called superconvergence [22]). See Table 1 for details about experimental settings.

Performance evaluation

We evaluated the models using the two different test sets described above: (i) data put aside from the training data repositories, 10% of each, making sure there were no subjects appearing in both training and test and controlling for age by stratification over age groups; (ii) the IXI dataset, i.e. data from a completely independent study of 561 subjects, simulating a more realistic use-case for the models. For the hold-out set in (i), we report both the overall results and the results on each repository.

As performance metrics we used the Sørensen-Dice similarity coefficient (DSC) and the Jaccard index (Jacc), measuring the degree of overlap between the ground truth masks generated by FSL and the model predictions. We also used the Hausdorff distance (Haus) between the two masks as a metric. The DSC is the mean overlap of the masks, while

Jacc is the union overlap, and Haus is a measure of extreme deviation between the masks:

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}, \qquad \text{Jacc} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}, \qquad \text{Haus} = \max\{(h(X,Y), h(Y,X)\}$$

where $h(X,Y)$ identifies the voxel $x \in X$ that is farthest from any voxel of Y and measures the distance from x to its nearest neighbor in Y. This means that $h(X,Y)$ first looks for the nearest voxel in Y for every voxel in X, and then the largest of these values are taken as the distance, which is the most mismatched point of X. Similarly for $h(Y,X)$, meaning that $\text{Haus}(X,Y)$ is able to measure the degree of mismatch between ground truth X and prediction Y from the distance of the point of X that is farthest from any point of Y, and vice versa.

A data filter

While looking at some training images and their corresponding ground truth labels, we observed a few images that were incorrectly labeled as shown in Fig. 3. In order to cope with this issue, we trained a model on the entire training set (training and validation) for a few epochs, and manually looked at the data having DSC < 0.8.

By applying this approach we ended up removing 14 images from the training set before training our final model. Note that we used the same Dice threshold to look at predictions made on test data and IXI with our final model, which led to removing additional 12 images (7 test + 5 IXI). Note also that all images that were removed were clear FSL failures, not prediction failures, confirmed by visual inspection.

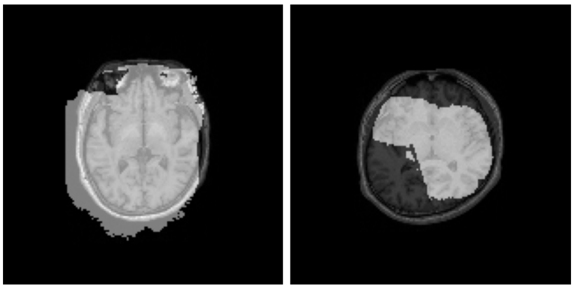

Figure 3: *Three instances of `fsl_anat` segmentation failure observed in our dataset.*

3 Results

Figure 4 depicts pair-wise 2D/3D comparative violin plots with jittering showing the distribution of the Dice coefficient for all MRI examinations across the collection of test data cohorts. From this, we observe close to negligible differences in performance between our 2D and 3D models. This is further illustrated by the results in Table 2, showing only small differences in the performance metrics.

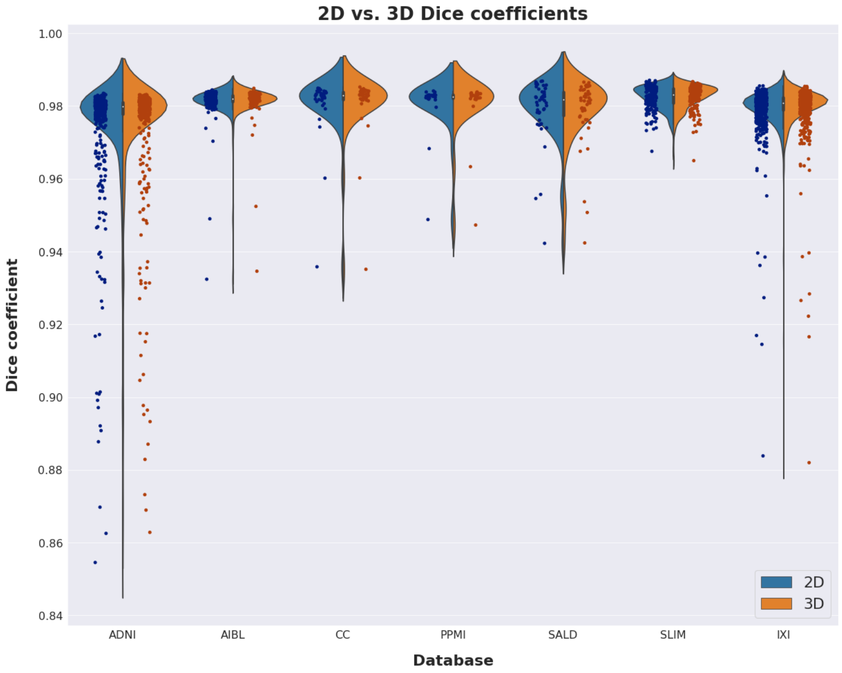

Figure 4: Violin plot of the Dice scores obtained by our models on the test dataset. Column names at the bottom of the plot refer to their database sources. The dots along the lower tail of the DSC distributions indicate outliers. A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure4`.

	Test			IXI		
	Dice	**Jaccard**	**Hausdorff**	**Dice**	**Jaccard**	**Hausdorff**
2D U-Net	0.9778 (0.0131)	0.9569 (0.024)	5.6711 (4.7215)	0.9791 (0.0076)	0.9591 (0.0140)	5.7811 (5.4826)
3D U-Net	0.9781 (0.0133)	0.9574 (0.024)	5.0558 (6.4009)	0.9796 (0.0077)	0.9601 (0.0140)	5.9220 (2.7330)

Table 2: The average (SD) values of Dice score, Jaccard Index and Hausdorff distance on the test datasets (Test and IXI) for our 2D U-Net and 3D U-Net models.

On a standard CPU, making predictions, including loading the image data into memory, on the two test datasets (test and IXI) took 1.38 ± 0.05 s and 1.37 ± 0.02 s for the 3D model and 12.36 ± 0.57 s and 11.82 ± 0.54 s for the 2D model[5].

4 Discussion

Using a large collection of T1w MR images sourced from a variety of openly available datasets and a well-established set of FSL tools for automated generation of “ground truth" brain masks, we have constructed 2D and 3D models for fast and accurate skull stripping. On independent test sets our models were able to produce brain masks that are very close to those produced by the much slower FSL-based process ($\sim$ 10 mins per volume), and even in some cases demonstrating higher robustness than the slower approach (Fig. 5).

[5] On a single GPU the time for inference for the 3D model was 0.59 ± 0.05 s and 0.57 ± 0.01 s on the two test datasets

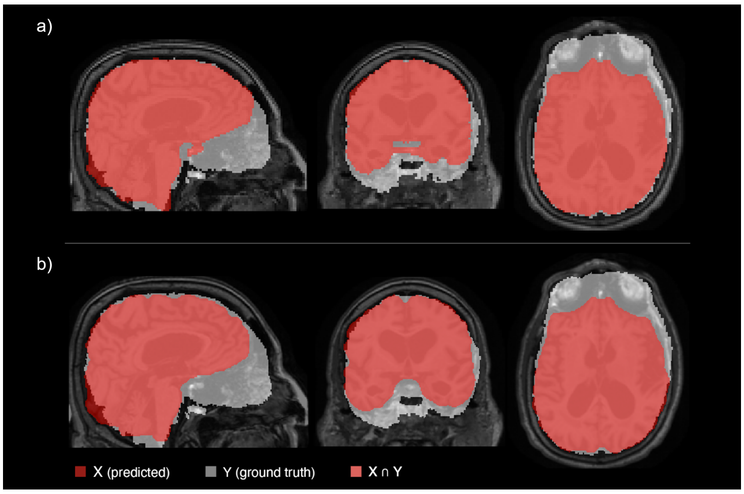

Figure 5: Comparison of (a) the brain mask produced by our 3D U-Net model on a dataset from ADNI achieving a poor Dice score, DSC < 0.9, and the corresponding ground truth FSL mask, and (b) the same comparison of the HD-BET model from [11]. Note the large amount of misclassification (extra-cerebral detection) of brain tissue by the “ground truth" FSL skull stripping procedure. This indicate that CNN-models may have some robustness advantages over FSL and perhaps also other traditional skull stripping . A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure5`.

In our comparisons between the 2D and 3D approaches we found similar performance as measured by Dice scores, Jaccard Index and Hausdorff distance, but also that the slice-by-slice based predictions necessary for the 2D approach made it significantly slower.

To decrease the variation in the training data images we performed bias-field correction before the images were fed to the network. This means that the networks have seen less bias than naturally occurs. To investigate the impact of this design decision we evaluated the trained models on the non-bias field corrected test images, reoriented to the standard MNI152 orientation, and also on and image with a high bias field shown in Fig. 2. Our 3D model had a Dice score of 0.978 ± 0.014 on the test set and 0.979 ± 0.008 on the IXI dataset when fed the uncorrected images. On the single high-bias field image displayed in Fig. 2, the model had a Dice score of 0.956 on the bias-field corrected image and a Dice score of 0.955 on the uncorrected image (Fig. 6).

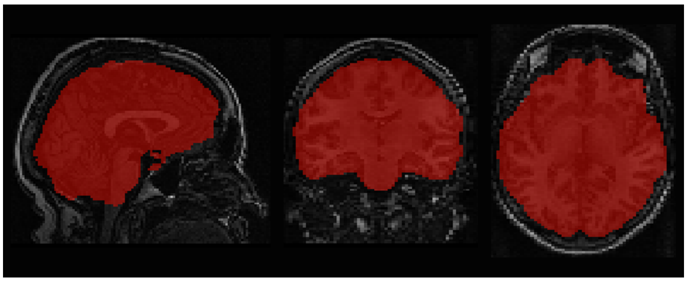

Figure 6: Predicition on a T1w image recorded at our own institution [5]. A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure6`.

Having a fast and accurate skull stripping method can have practical utility as it can speed up larger image processing pipelines, e.g. for subcortical segmentation or segmentation of other regions of interest like brain tumors or lesions. Once the accuracy reaches a certain threshold, issues related to defining the ground truth becomes more important than increased accuracy at reproducing said ground truth labels. This can be illustrated by the different labels used in our work and in the HD-BET work of [11] described above. Feeding our test images through the trained HD-BET model results in an average Dice score of 0.9615 ± 0.0295. This does not mean that their model performs worse than ours at skull stripping, only that the ground truth labels used when training the models differs. Robustness also becomes more important than increasing the accuracy. As indicated in Fig. 5, CNN-based models may have an advantage here, but this requires further investigation.

Using our approach in a setting with various pathologies will require further investigations of its robustness to such variation in the images, and also clarification of “ground truth" consensus. Training the models on datasets that includes images with pathologies, and also adding an automized MRI quality control system based on e.g. MRIQC [23], would be natural next steps.

For thorough validation in a realistic setting, embedding the models in established workflows is key. At our hospital we have recently established a PACS, RIS and EDC system for research that integrates with the clinical systems. This enables real-world testing of this and other image processing methods, a crucial step for bringing deep learning research into practice [24].

5 Acknowledgments

Our work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07.

Data collection and sharing for this project was partly funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] A. S. Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on MRI,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2.

[2] E. H. Shortliffe and B. G. Buchanan, “A model of inexact reasoning in medicine,” *Mathematical biosciences*, vol. 23, no. 3-4, pp. 351–379, 1975.

[3] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[4] T. Ching *et al.*, “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.

[5] M. A. Ystad, A. J. Lundervold, E. Wehling, T. Espeseth, H. Rootwelt, L. T. Westlye, M. Andersson, S. Adolfsdottir, J. T. Geitung, A. M. Fjell, I. Reinvang, and A. Lundervold, "Hippocampal volumes are important predictors for memory function in elderly women," *BMC Med Imaging*, vol. 9, no. 1, 2009.

[6] P. Kalavathi and V. S. Prasath, "Methods on skull stripping of MRI head scan images – a review," *Journal of Digital Imaging*, vol. 29, no. 3, pp. 365–379, 2016.

[7] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, 2002.

[8] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL," *NeuroImage*, vol. 45, no. 1, pp. S173–S186, 2009.

[9] B. Avants, A. Klein, N. Tustison, J. Woo, and J. C. Gee, "Evaluation of open-access, automated brain extraction methods on multi-site multi-disorder data," in *16th annual meeting for the Organization of Human Brain Mapping*, 2010.

[10] R. W. Cox, "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical research*, vol. 29, no. 3, pp. 162–173, 1996.

[11] F. Isensee *et al.*, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human brain mapping*, vol. 40, no. 17, pp. 4952–4964, 2019.

[12] R. Dey and Y. Hong, "CompNet: Complementary segmentation network for brain MRI extraction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 628–636, Springer, 2018.

[13] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

[14] G. Gavidia-Bovadilla, S. Kanaan-Izquierdo, M. Mataró-Serrat, A. Perera-Lluna, A. D. N. Initiative, *et al.*, "Early prediction of Alzheimer's disease using null longitudinal model-based classifiers," *PloS one*, vol. 12, no. 1, p. e0168011, 2017.

[15] K. A. Ellis *et al.*, "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease," *International psychogeriatrics*, vol. 21, no. 4, pp. 672–687, 2009.

[16] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: DICOM to NIfTI conversion," *Journal of neuroscience methods*, vol. 264, pp. 47–56, 2016.

[17] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[18] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, 2001.

[19] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, 2002.

[20] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[21] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *arXiv preprint arXiv:1707.02937*, 2017.

[22] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, International Society for Optics and Photonics, 2019.

[23] O. Esteban, D. Birman, M. Schaer, O. O. Koyejo, R. A. Poldrack, and K. J. Gorgolewski, "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites," *PloS one*, vol. 12, no. 9, 2017.

[24] M. Nagendran *et al.*, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, 2020.