



**TECHNICAL REPORT**

# Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks

Maged Goubran<sup>1,2</sup>  | Emmanuel Edward Ntiri<sup>1,2</sup> | Hassan Akhavein<sup>1,2</sup> |  
 Melissa Holmes<sup>1,2</sup> | Sean Nestor<sup>1,3</sup> | Joel Ramirez<sup>1,2</sup> | Sabrina Adamo<sup>1,2</sup> |  
 Miracle Ozzoude<sup>1,2</sup> | Christopher Scott<sup>1,2</sup> | Fuqiang Gao<sup>1,2</sup> | Anne Martel<sup>4</sup> |  
 Walter Swardfager<sup>2,5</sup> | Mario Masellis<sup>2,6</sup> | Richard Swartz<sup>1,2,6</sup> |  
 Bradley MacIntosh<sup>2,4</sup>  | Sandra E. Black<sup>1,2,7</sup>

<sup>1</sup>LC Campbell Cognitive Neurology Unit, Hurvitz Brain Sciences Research Program, Sunnybrook Research Institute, University of Toronto, Toronto, Ontario, Canada

<sup>2</sup>Canadian Partnership for Stroke Recovery, Heart and Stroke Foundation, Toronto, Ontario, Canada

<sup>3</sup>Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada

<sup>4</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

<sup>5</sup>Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada

<sup>6</sup>Department of Medicine (Neurology division), University of Toronto, Toronto, Ontario, Canada

<sup>7</sup>Department of Medical Imaging, University of Toronto, Toronto, Ontario, Canada

**Correspondence**

Maged Goubran, 2075 Bayview Avenue, M6  
 West RM 176, Toronto, ON M4N 3M5,  
 Canada.

Email: maged.goubran@sri.utoronto.ca

**Funding information**

Canadian Institute for Health Research (CIHR) MOP Grant, Grant/Award Number: 13129; CIHR Foundation, Grant/Award Number: 159910; L.C. Campbell Foundation; Alzheimer's Disease Neuroimaging Initiative (ADNI), Grant/Award Number: U01 AG024904; Department of Defense ADNI, Grant/Award Number: W81XWH-12-2-0012; National Institute on Aging; National Institute of Biomedical Imaging and Bioengineering; 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 and Development LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx

**Abstract**

Hippocampal volumetry is a critical biomarker of aging and dementia, and it is widely used as a predictor of cognitive performance; however, automated hippocampal segmentation methods are limited because the algorithms are (a) not publicly available, (b) subject to error with significant brain atrophy, cerebrovascular disease and lesions, and/or (c) computationally expensive or require parameter tuning. In this study, we trained a 3D convolutional neural network using 259 bilateral manually delineated segmentations collected from three studies, acquired at multiple sites on different scanners with variable protocols. Our training dataset consisted of elderly cases difficult to segment due to extensive atrophy, vascular disease, and lesions. Our algorithm, (HippMapp3r), was validated against four other publicly available state-of-the-art techniques (HippoDeep, FreeSurfer, SBHV, volBrain, and FIRST). HippMapp3r outperformed the other techniques on all three metrics, generating an average Dice of 0.89 and a correlation coefficient of 0.95. It was two orders of magnitude faster than some of the tested techniques. Further validation was performed on 200 subjects from two other disease populations (frontotemporal dementia and vascular cognitive impairment), highlighting our method's low outlier rate. We finally tested the methods on real and simulated "clinical adversarial" cases to study their robustness to corrupt, low-quality scans. The pipeline and models are available at: <https://hippmapp3r.readthedocs.io> to facilitate the study of the hippocampus in large multisite studies.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2019 The Authors. *Human Brain Mapping* published by Wiley Periodicals, Inc.

Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; Transition Therapeutics

## KEYWORDS

brain atrophy, convolutional neural networks, deep learning, dementia, hippocampus, image segmentation

## 1 | INTRODUCTION

The hippocampus is implicated in many neurological diseases including Alzheimer's disease (AD), epilepsy, and schizophrenia, among others (Barnes et al., 2009; Bernasconi, Natsume, & Bernasconi, 2005; Goubran et al., 2016; Santyr et al., 2017; Steen, Mull, McClure, Hamer, & Lieberman, 2006). Hippocampal volumetry has been found to be a critical biomarker of aging and dementia (Courchesne et al., 2000; Jack Jr et al., 1998). It represents a central correlate of memory function and is widely used as a predictor of cognitive decline, both in research and clinical settings (Jack et al., 2000; Rusinek et al., 2003; Schill et al., 2003; Sullivan, 2002). Hippocampal delineation is also employed for investigation of the hippocampal-neocortical connectivity and studying diffusion magnetic resonance imaging (MRI) changes in the medial temporal lobe. The hippocampal anatomy is variable, and its complex structure is selectively affected in different disorders (Goubran et al., 2014). Manual segmentation of the hippocampus is very time consuming and may suffer in reproducibility across different raters. While numerous algorithms have been developed for automated segmentation of the whole hippocampus (Fischl et al., 2002; Iglesias et al., 2015; Nestor et al., 2013; Thyreau, Sato, Fukuda, & Taki, 2018), the overwhelming majority suffer from at least one the following issues: (a) the algorithms are not made publicly available or the trained models are not released, (b) they have been trained on young adult brain images and are unable to accurately deal with brain atrophy or lesions associated with aging and neurodegeneration, or (c) they require parameter tuning, large computational time or advanced programming knowledge to execute them. With the increasing amounts of data and large multicenter studies, there is a great need for efficient, easy-to-use software that performs accurate quantification of structural biomarkers in elderly subjects while also enabling personalized assessments.

Recently, deep neural networks, and particularly convolutional neural networks (CNNs), have shown superior performance to other machine learning techniques on computer vision tasks such as image classification (Krizhevsky, Sutskever, & Hinton, 2012) and semantic segmentation (Long, Shelhamer, & Darrell, 2015). These deep networks have been more recently applied in medical imaging (Çiçek, Abdulkadir, Lienkamp, Brox, & Ronneberger, 2016; Kamnitsas et al., 2017; Kayalibay, Jensen, & van der Smagt, 2017; Milletari, Navab, & Ahmadi, 2016; Ronneberger, Fischer, & Brox, 2015), among other domains. However, there are several challenges to applying these networks to biomedical data. These supervised machine learning techniques, specifically deep networks, require very large amounts

of labeled (ground truth) data, typically in the thousands or millions in the computer vision field, in order to train and optimize millions of weights. Creating databases of manually delineated ground truth labels (in 3D) for medical images requires a large amount of time and training, and hence these databases are scarce or commonly consist of smaller cohorts, typically in the hundreds or even less. Most of the networks developed for computer vision applications rely on a 2D architecture which is suitable for stacks of 2D images or 3D images with small depth. Most whole brain T1-weighted structural brain scans have close to isotropic resolutions, that is, similar sizes in each dimension, making slice-by-slice application of 2D architectures inefficient (Kayalibay et al., 2017). Novel network architectures tend to train on larger sections or entire images as opposed to small patches. This training approach creates a class imbalance as it is bound to the original distribution of classes in the dataset, which in medical images is dominated by background (negative) voxels (Kamnitsas et al., 2017).

In this paper, we present HippMapp3r, an open-source, efficient whole hippocampal segmentation algorithm based on 3D CNNs that is robust to brain atrophy due to neurodegenerative changes. Our deep learning-based segmentation model was trained on a large database consisting of 209 meticulously hand-drawn segmentations of elderly subjects with brain atrophy and lesions from multiple studies. Individuals in the current cohort spanned a range of cognitive neurology presentations: cognitively unimpaired controls, patients with mild cognitive impairment (MCI), AD, or temporal lobe epilepsy (TLE). Consequently, numerous types of brain injury were included in the datasets: gray matter atrophy, ventricular enlargement, white matter hyperintensities (WMH) and perivascular spaces. These scans were part of multisite studies using different scanners, field strengths, and scanning protocols. We built 3D CNNs with a U-net architecture, residual units, and a weighted dice coefficient loss function to deal with class imbalance. The developed model was validated against state-of-the-art techniques to highlight its accuracy and efficiency. We also investigated outliers and failure rates in all tested methods using two additional patient populations, frontotemporal dementia (FTD) and vascular cognitive impairment (VCI). We further tested our model on corrupt data that did not pass quality control and on simulated realistic (clinical) adversarial attacks through sharp decreases in resolution, signal-to-noise ratios (SNR), and cropping of field of view (FOV). We are making our tools and models available to the research community and developed an easy-to-use pipeline with a graphical user interface (GUI) and thorough documentation to make it accessible to users without extensive programming knowledge.

## 2 | METHODS

### 2.1 | Participants and image acquisition

A total of 259 participants were used for the hippocampal segmentation model, combined from three separate studies: 100 were recruited from the Sunnybrook Dementia Study (SDS) (Deshpande et al., 2004), 135 from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and the harmonization study (Boccardi et al., 2015) and 24 from the University of Pennsylvania (UPenn) TLE atlas (Das et al., 2009). Table 1 presents left and right hippocampal volumes, as well as WMH and stroke volumes to describes the ranges of hippocampal atrophy and lesions in these cohorts.

All SDS patients were recruited from the LC Campbell Cognitive Neurology Research Unit, Sunnybrook Health Sciences Centre at the University of Toronto (age:  $71 \pm 10$ , 28 Males). The T1-weighted three-dimensional volumetric scan was acquired using a 1.5 Tesla Signa system (GE Healthcare, Chicago, Illinois). The acquisition parameters for the T1 spoiled gradient echo sequence were: 124 slices; matrix,  $256 \times 192$ ;  $22 \times 16.5$  cm FOV; number of excitations, 1; echo time/repetition time, 35 ms/5 ms; flip angle,  $35^\circ$ , and in-plane resolution of  $0.859 \times 0.859$  mm with slice thickness between 1.2 and 1.4 mm depending on head size. ADNI subjects were scanned at multiple sites using a mixture of 1.5 T ( $N = 68$ ) and 3 T ( $N = 67$ ) scanners (age:  $75 \pm 8$ , 70 males). Participants were nearly equal in cohort size for the three diagnostic groups (normal, MCI and AD), and three scanner major manufacturers (GE, Siemens, and Philips). The ADNI-GO/2 MRI protocol has been optimized to provide comparable images from different 3 T platforms from the three manufacturers. The T1 weighted magnetization prepared rapid gradient echo (MPRAGE) had the following parameters: TR/TE/TI = 2300/2.95/900 ms, sagittal,  $1.1 \times 1.1 \times 1.2$  mm spatial resolution. UPenn subjects were scanned at a 3 T Siemens Trio scanner using an eight-channel head coil and body coil transmitter. The T1-weighted structural MRI scan used the MPRAGE sequence with the following parameters: TR = 1,620 ms, TE = 3.87 ms, TI = 950 ms, flip angle =  $15^\circ$ , and voxel size  $0.9375 \times 0.9375 \times 1$  mm.

### 2.2 | Model architecture and contributions

Our algorithm consists of a serial "ensemble" of two networks, an initial network trained on the whole brain and a second network with the same architecture trained on the first network's output (operating on a reduced FOV centered around the initial segmentation). Defining the architecture of deep CNN networks and loss functions are important factors in the construction of a deep model and are often guided by the specific application the network is set to achieve. Our CNN networks are based on a convolutional autoencoder-like (U-net) architecture (Çiçek et al., 2016; Milletari et al., 2016; Ronneberger et al., 2015), which consists of contracting and expanding pathways (stages) and is trained on the entire image rather than patches. The contracting pathway encodes context (representations of the input) and the expanding pathway recombines

**TABLE 1** Participants study demographics, clinical diagnosis, MMSE scores, WMH and stroke volumes, and MRI field strength in the train and test datasets

Population	N	Age	Sex	Dx	R Hp volume (mm <sup>3</sup> )	L Hp volume (mm <sup>3</sup> )	WMH volume (cc)	Stroke volume (cc)	MMSE	Field strength	GT	
Train (N = 209)	SDS	80	85.3 ± 11.2	56% M/44% F	58% AD, 26% NC, 16% VCI	2,700.8 ± 537.5 (1,610.8, 3,815.6)	2,633.57 ± 553.82 (1,417.2, 3,854.9)	11.37 ± 16.47	0	21.41 ± 15.12	1.5 T	Y
	ADNI	109	74.1 ± 7.8	51% M/49% F	43% AD, 16% NC, 10% MCI	2,797.4 ± 582.32 (1,054.0, 5,029.0)	2,693.57 ± 586.28 (1,428.1, 5,140.0)	1.71 ± 3.95	0	22.02 ± 6.99	1.5 T/3 T (50/50%)	Y
	Upenn	20			100% TLE	3,224.4 ± 482.9 (2096.6, 4,948.6)	3,124.6 ± 482.5 (2059.4, 4,948.6)	—	—	—	3 T	Y
Test (N = 300)	Mixture from train set	50									1.5 T/3 T	Y
	FTLD	95	78.0 ± 11.4	52% M/48% F	100% FTLD	3,039.6 ± 591.3 (1,470.3, 4,061.6)	2,813.6 ± 563.9 (1,321.1, 3,938.7)	5.49 ± 9.36	0	22.27 ± 7.11	1.5 T	N
	VCI	105	90.6 ± 9.5	54% M/46% F	100% VCI	3,043.6 ± 618.2 (494.5, 4,385)	2,898.8 ± 634.7 (558.8, 4,515)	18.86 ± 19.93	17.14 ± 34.15	23.95 ± 4.58	1.5 T	N
	Adversarial cases	50	90.5 ± 5.6	64% M/36% F	44% AD, 24% NC, 18% VCI, 14% FTD	3,173.6 ± 620.1 (1,772.9, 4,457.0)	3,050.5 ± 639.0 (1,282.0, 4,040.0)	—	—	21.38 ± 6.34	1.5 T	N

Abbreviations: AD, Alzheimer's disease; ADNI, Alzheimer's disease neuroimaging initiative; Dx, diagnosis; FTLD, fronto-temporal lobar dementia; GT, ground truth; Hp, hippocampus; MMSE, mini-mental state examination; MRI, magnetic resonance imaging; NC, normal controls; SDS, Sunnybrook Dementia Study; Upenn, University of Pennsylvania; VCI, vascular cognitive impairment; WMH, white matter hyperintensity.

encoded representations with shallower features to enable precise localization of the voxels of interest (i.e., the hippocampus). The overall architecture of the proposed 3D network (Figure 1) is inspired by the original 2D U-net (Ronneberger et al., 2015) with a few modifications: (a) In this work we updated the original design with residual blocks (He, Zhang, Ren, & Sun, 2016; Kayalibay et al., 2017; Milletari et al., 2016) that ease optimization convergence by improving gradient flow and enable higher accuracy through a deeper network. (b) Our custom residual blocks consisted of two convolution blocks (convolution layer with normalization and nonlinearities) separated by a dropout layer to avoid overfitting. (c) We employed deep supervision (Kayalibay et al., 2017; Lee, Xie, Gallagher, Zhang, & Tu, 2014) at the expanding pathway by adding earlier feature maps at different levels of the network and combining them via element-wise summation to form the final network output. (d) Since the 3D network is memory expensive, we opted to use instance normalization (Iseensee, Kickingereder, Wick, Bendszus, & Maier-Hein, 2018; Ulyanov, Vedaldi, & Lempitsky, 2016) instead of the commonly used batch normalization as the stochasticity generated by a small batch size may destabilize batch normalization. (e) To generate segmentation maps from the entire input image, we relied on trainable deconvolution kernels as the upsampling operations. (f) Finally, we chose a loss function based on the Dice similarity coefficient (Dice, 1945) (see Section 2.3).

The network is fully convolutional, that is no fully connected layers are added, and hence can predict a variable number of voxels in a forward pass without the need for architectural changes (Long et al., 2015). It employs skip connections to combine feature maps across stages through concatenation. The network has a depth of five and 16 initial filters, with the number of filters doubling every contraction step. The building blocks of the networks are convolution blocks, consisting of a convolution layer followed by a normalization layer and a

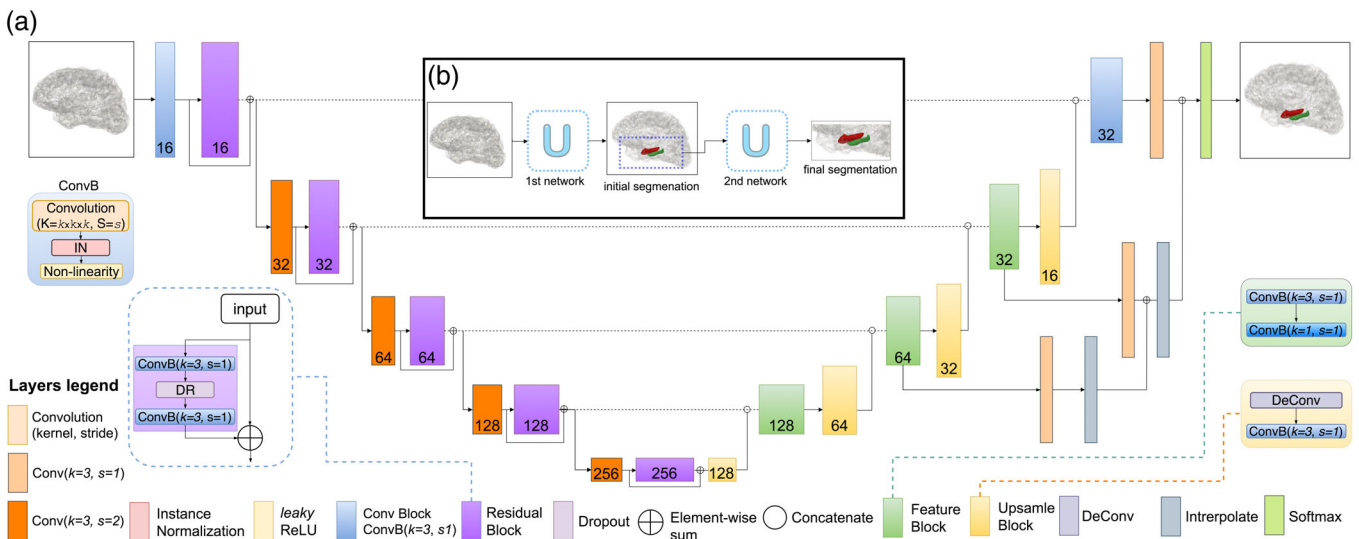
nonlinearity. We chose the leaky ReLU as the activation function with a negative slope of  $10^{-2}$  for all feature map convolutions.

The residual blocks separate the input data into two paths, the first applies weights and nonlinearities, and the second applies an identity mapping (the input data is unchanged). The two paths are finally merged with the element-wise sum, resulting in the following formulation:

$$y(x) = \sigma(W_2\sigma(W_1x) + x) \tag{1}$$

where  $W_1$  and  $W_2$  are the weights of the convolutional layers and  $\sigma$  is the activation function. The residual block consisted of two convolution blocks, separated by a dropout layer. We chose a dropout rate of 0.3 and a small convolution kernel of  $3 \times 3 \times 3$ , which enables a higher nonlinearity capacity for the same receptive field with a lower number of parameters needed. The residual blocks are preceded by a strided convolution layer with a kernel of  $3 \times 3 \times 3$  and a stride of two, effectively reducing the number of feature dimensions by a factor of two while adding more features as the network depth increases.

In the expanding pathway, we relied on upsampling blocks that repeat the feature voxels twice in each spatial dimension, followed by concatenation to combine the upsampled features with those from the corresponding contracting pathway. After concatenation, a feature block recombines these features together and halves the number of feature maps at each step to reduce memory consumption. Upsampling blocks consisted of deconvolution layers, followed by a convolution block with kernel  $3 \times 3 \times 3$ . As for the feature blocks, they consisted of two consecutive convolution blocks with the first having a convolution layer with a kernel size of  $3 \times 3 \times 3$  and the second having a kernel of  $1 \times 1 \times 1$ . As discussed before, we added



**FIGURE 1** Proposed base network architecture with insets for the convolution, residual, feature and upsample blocks. (a) The 3D layers are color-coded to distinguish their different functionality and presented as 2D representations for simplification. For each layer, the number of features is shown at the bottom. “Conv” represents a convolution, “k” represents the kernel, “s” denotes the number of strides. (b) Overall scheme demonstrating the use of an ensemble of two consecutive networks to produce the final segmentation

earlier feature maps at different levels ( $n = 3$ ) of the network and combined them via element-wise summation to form the final output. Feature maps from the last layer were passed to a softmax function that generates pseudo-class probability maps as:

$$\rho_c(H_L) = \frac{\exp(H_L)}{\sum_{c=1}^C \exp(H_L)} \forall c, \quad (2)$$

where  $c$  denotes class and  $C$  the total number of classes, that is, either a hippocampus voxel or not, in-turn producing a class-likelihood probability for each voxel in the image.

### 2.3 | Loss function

To mitigate the class imbalance issue (the majority of image voxels do not represent the structure of interest), previous work (Çiçek et al., 2016; Ronneberger et al., 2015) applied a weight map to the categorical cross-entropy loss (objective) function. We sought to maximize the Dice similarity coefficient (Dice, 1945) as suggested by (Milletari et al., 2016). We employed a formulation of Dice similarity as an equally weighted dice loss function. An advantage of this approach is that it does not rely on hyperparameters. The Dice similarity (Dice, 1945) is an overlap metric commonly used to quantify segmentation accuracy. It is defined as follows between two binary volumes:

$$D = \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \quad (3)$$

where  $p_i$  are voxels of the predicted binary segmentation volume  $p_i$ ?  $P$  and  $g_i$  are voxels of the ground truth volume  $g_i$ ?  $G$ . This formulation is differentiable and can be incorporated in the network, yielding the following gradient:

$$\frac{\partial D}{\partial p_j} = 2 \left[ \frac{g_j \left[ \sum_i^N p_i^2 + \sum_i^N g_i^2 \right] - 2 p_j \left[ \sum_i^N p_i g_i \right]}{\left[ \sum_i^N p_i^2 + \sum_i^N g_i^2 \right]^2} \right] \quad (4)$$

### 2.4 | Data preprocessing, augmentation and model training

Prior to training, all images were bias field corrected for B1 inhomogeneities using N4 (Tustison et al., 2010). They were then standardized to have a zero mean and unit variance within a local neighborhood of 50 voxels using  $c3d$  (Yushkevich et al., 2006). We opted for neighborhood normalization instead of global image normalization to better preserve local features. We performed a total of three augmentations per scan including flipping images along Left-Right and  $\pm 15^\circ$  rotations in the L-R axis. SDS datasets were not L-R flipped as the data comes from 50 unique subjects segmented twice each (after L-R flipping), generating 100 ground truth segmentations. No deformable augmentation was employed, relatively preserving the anatomy of

input images. Models were trained for 300 epochs and early stopping was set to 50 epochs where validation loss did not improve. We used the Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of  $5 \times 10^{-3}$ , a patience of 10 epochs for the validation loss and a learning rate drop (decay factor) of 0.5.

Both networks trained on single contrast T1-weighted image inputs. The initial 3D U-net trained on downsampled T1 images (to a size of  $128 \times 128 \times 128$ ), and the second network trained on a limited FOV of  $112 \times 112 \times 64$  voxels centered around the initial segmentation. The initial network was trained on a mixture of skull-stripped and intact-skull images. Each network was trained with five-fold cross validation. Out of the 259 datasets with manual hippocampal segmentations used in this study, 184 (~70%) were used for training, 50 (~20%) for testing and 25 (~10%) for validation during training. With data augmentation, this split generated a total of 502 training samples. The networks were implemented using Keras (using Tensorflow backend) and trained on a GeForce GTX1080 Ti graphics card with 11Gb of memory and a Pascal architecture (NVIDIA, Santa Clara, CA). The algorithm and trained model are available at: <https://hippomap3r.readthedocs.io>.

### 2.5 | Evaluation of clinical datasets

The model was tested on two datasets (Table 1). Fifty subjects with manually traced ground truth from the aforementioned studies were used for the first dataset. Images from 200 additional subjects participating in the SDS were used in the second dataset from two separate disease cohorts: FTD who have severe atrophy particularly in the temporal regions, and VCI who typically have strokes and severe WMH burden. The disease cohorts in this test set are characterized by atrophy, vascular lesions and features not present in our training set. The second set was employed to investigate outliers and failure rates. Manual ground truth segmentation was not available for this set.

The model was compared against established state-of-the-art techniques for the two test sets: FreeSurfer's whole hippocampus segmentation (version 6.0) (Fischl et al., 2002), FSL's subcortical segmentation (FIRST) (v. 5.0.10) (Patenaude, Smith, Kennedy, & Jenkinson, 2011), an in-house developed segmentation tool (SBHV) (v. 1.0) (Nestor et al., 2013), a multi-atlas patch-based model (volBrain) (v. 1.0) (Manjón & Coupé, 2016), and a newly developed CNN-based model (Hippodeep) (v. 0.1) (Thyreau et al., 2018). FreeSurfer's hippocampus algorithm combines in-vivo and ex-vivo tracings into a computational atlas that employs Bayesian inference to segment the hippocampal subfields. FIRST is a Bayesian model-based tool with deformable surfaces that rely on shape and appearance for segmentation. SBHV employs a multi-atlas-based segmentation approach. Hippodeep uses a relatively shallow CNN trained on both FreeSurfer's hippocampal segmentation from multiple large cohorts and augmented iterations of a much smaller manually ground-truth dataset. volBrain is an open-source, automatic online tool that provides a series of image segmentation tasks. It employs a modified edition of nonlocal label fusion for subcortical structure segmentation.

While there were slight differences in protocols between the tested segmentation methods and the ground truth used for our models, all protocols used very similar border definitions for the hippocampus. ADNI, volBrain, and SBHV used the EADC-ADNI Harmonized Protocol (HarP) (Boccardi et al., 2015), consisting of the hippocampal head, body, the alveus/fimbria up to the separation from the fornix, the medial border of the hippocampal body and subiculum, and the whole hippocampal tail. The UPenn atlas used data that was segmented using a semiautomated pipeline (Pluta et al., 2009) and included the hippocampus proper, the dentate gyrus, the alveus, the fimbria, and the subiculum (Anon, 2005; Hasboun et al., 1996). Like the previous methods, the tracings included all rostrocaudal parts of the hippocampus. FIRST, Hippodeep, and Freesurfer share the same protocol for segmenting the hippocampus, which included the dentate gyrus, the hippocampus proper, the prosubiculum, and the subiculum (<http://freesurfer.net/fswiki/CMA>). Careful distinction is paid between the hippocampus, the amygdala, the thalamus and the temporal horn of the lateral ventricle (Schoemaker et al., 2016).

## 2.6 | Clinical adversarial attacks

To further validate our model, we tested it on challenging images ( $n = 50$ ) that did not pass our quality control (QC) protocol and were deemed corrupt due to motion, low SNR and ringing artifacts; herein referred to as “adversarial attacks,” a commonly used term in the deep learning field referring to engineered inputs with perturbations presented to neural networks in order to drive them to produce errors and study their robustness toward different inputs.

Furthermore, we performed additional validation experiments whereby we simulated low-quality clinical grade or challenging scans acquired with different acquisitions and scanners. These simulated adversarial attacks included: (a) decreases in resolution, (b) addition of noise, and (c) cropping the FOV. Input images were downsampled by a factor of two in all dimensions, and 2× in-plane with 4× out-of-plane (in the z-dimension), to simulate clinical imaging protocols which tend to acquire images with lower resolutions and thicker slices. Speckle noise with a  $\sigma = \{0.1, 0.3\}$  was applied to input images to simulate scans with lower SNR and/or low-field strength acquisitions. Salt and pepper noise with a  $\sigma = 0.1$  were also applied to simulate signal drop-out in random voxels. Finally, we also cropped the input sequences in the Superior–Inferior plane 15% from each side to simulate scans with a limited FOV.

## 2.7 | Validation metrics

Four metrics were used to evaluate model performance against manual segmentation: the Pearson R correlation coefficient of the volumes, the Dice similarity coefficient, the Jaccard coefficient, and the Hausdorff distance. We used the Pearson R correlation coefficient (Pearson, 1895) between manually segmented volumes and volumes generated through model predictions to assess the clinical utility of the predictions. The Dice coefficient has been described in Section 2.3 and the Jaccard coefficient (Jaccard, 1912) is another

metric to assess the degree of similarity between two sets (sometimes referred to as the intersection over the union or IoU) as defined by:

$$Jacc = \frac{|P \cap G|}{|P \cup G|} \quad (5)$$

where  $P$  and  $G$  are the predicted and ground truth masks. The Hausdorff distance was used to evaluate the similarity in shape between the ground truth and each segmentation method. The Hausdorff distance of two objects in the same space is defined as the largest distance in a set of all closest distances between both sets of points:

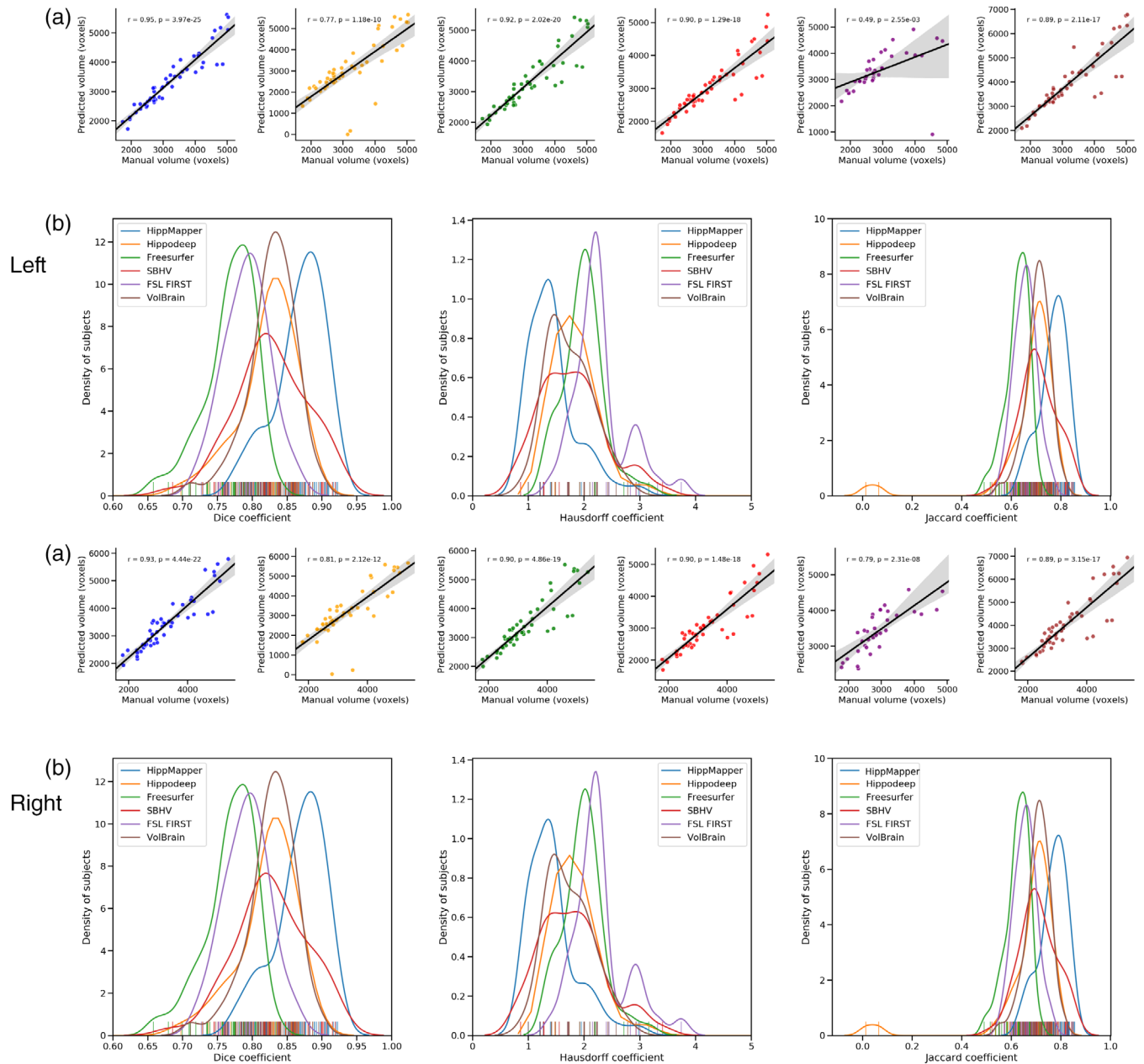
$$H(A, B) = \max\{\max_{x \in A}\{\min_{y \in B}\{\|x, y\|\}\}, \max_{y \in B}\{\min_{x \in A}\{\|x, y\|\}\}\}$$

Volumes and segmentations for both hemispheres were used for the quantitative analyses across all models. In the second test set, since no manual ground truth labels were available, we relied on computing Z-scores of the hippocampal surface areas of the subjects to highlight outliers and failed segmentations. For every subject, a mean surface area was computed from the different tested methods while excluding any values lower than the lower quartile (25th percentile) or greater than the upper quartile (75th percentile) by more than 1.5× the interquartile range ( $iqr = \text{upper quartile} - \text{lower quartile}$ ). The results were visually checked to further exclude any segmentations that did not pass quality control for quantification of subject means. Surface area z-scores were then computed for every method based on the filtered subject mean per participant, for each hemisphere separately. Very high or low z-score values then highlight potential outliers or failure cases (substantial deviation from the filtered mean computed across all methods).

## 3 | RESULTS

### 3.1 | Evaluation of clinical datasets

While all tested techniques provided significant correlations between predicted and manual volumes, our model generated the highest agreement with ground truth labels ( $r = 0.95$ ) and the lowest number of outliers (Figure 2a). The distributions of Dice and Jaccard coefficients between ground truth manual labels of the hippocampus and predicted segmentations are presented in Figure 2b. The shape of the distributions was computed using kernel density estimation with Gaussian (normal) kernels. Our network had a tight Dice coefficient distribution (similarly with Jaccard) centered around  $0.869 \pm 0.033$  ( $0.870 \pm 0.030$  left,  $0.868 \pm 0.036$  right) followed by SBHV, volBrain, FIRST, Hippodeep, and FreeSurfer. The SBHV pipeline had a better dice and standard deviation than volBrain, FIRST, Hippodeep, and FreeSurfer but with a trade-off of the longest computational time, taking at least 8 hrs per subject. While volBrain had a better Dice, Jaccard and Hausdorff distance than FIRST, Hippodeep and Freesurfer, the Dice accuracy for this patch-based algorithm may be lower than those reported in the literature due to the higher atrophy



**FIGURE 2** Validation of hippocampal segmentation through volume correlations, Hausdorff distances and Dice, Jaccard similarity coefficients on our proposed model, HippMap3r (blue) and five established techniques: Hippodeep (yellow), FreeSurfer (green), SBHV (red), FIRST (purple), and volBrain (brown). The proposed model produced the best agreement to manual labels among the six tested techniques in all four metrics. Results for all four metrics displayed for both left and right hippocampi. (a) Correlations between the manually segmented volumes and volumes generated through model predictions. Pearson R correlation coefficients and P-values are shown for each test. Our method produced the highest volume correlation for both hippocampi ( $r = .95, p < .000001$ ). (b) Distribution of Dice coefficients, Hausdorff distances, and Jaccard coefficients (in said order) between ground truth manual labels of the hippocampus and predicted segmentations. Ticks on the x-axis represent individual segmentation cases (colored by tested technique). Distributions are generated using Gaussian kernel density estimation

level in our test set (Manjón & Coupé, 2016). While Hippodeep and FreeSurfer's distributions were centered around 0.76 and 0.74, respectively, they had multiple cases with a Dice lower than 0.65. FIRST failed to run on eight subjects, producing errors without completion (mainly SDS cases) and generated misregistrations for another four, producing outputs with incorrect orientations. Hippodeep's standard deviation was more than 3x higher than the average standard

deviation of the other tested algorithms, possibly highlighting the instability of deep networks when presented with different inputs than the training set. Dice, Jaccard similarity coefficients, and Hausdorff distance between ground truth manual labels of the hippocampus and predicted segmentations for the six tested techniques are summarized in Table 2. Our model outperformed the other state-of-the-art techniques by 5% or more.

**TABLE 2** Dice, Jaccard coefficients, Hausdorff distances, and computational time for hippocampal segmentation methods

Hemisphere		HippMapp3r (mean ± std)	Hippodeep (mean ± std)	FreeSurfer (mean ± std)	SBHV (mean ± std)	FIRST <sup>a</sup> (mean ± std)	VolBrain (mean ± std)
Left	Dice coefficient	0.870 ± 0.030	0.781 ± 0.165	0.761 ± 0.031	0.832 ± 0.045	0.794 ± 0.028	0.825 ± 0.036
	Jaccard coefficient	0.771 ± 0.047	0.689 ± 0.073	0.615 ± 0.040	0.827 ± 0.051	0.645 ± 0.111	0.704 ± 0.050
	Hausdorff distance (mm)	1.487 ± 0.485	3.454 ± 6.296	2.172 ± 0.314	1.676 ± 0.584	2.798 ± 3.689	1.794 ± 0.463
Right	Dice coefficient	0.868 ± 0.036	0.775 ± 0.192	0.769 ± 0.037	0.827 ± 0.051	0.788 ± 0.036	0.828 ± 0.035
	Jaccard coefficient	0.768 ± 0.056	0.700 ± 0.057	0.625 ± 0.047	0.708 ± 0.074	0.657 ± 0.043	0.708 ± 0.050
	Hausdorff distance (mm)	1.440 ± 0.047	2.915 ± 5.281	1.988 ± 0.364	1.793 ± 0.594	2.291 ± 0.489	1.740 ± 0.407
	Approx. compute time	14 s	30 s	Whole brain: 6 hrs, Hipp.: 7 min (12 cores)	7 hrs	6 min	10 min

<sup>a</sup>FIRST failed to run (producing errors) on eight subjects and generated outputs with wrong orientation on another four.

To test whether an ensemble of networks would lead to a significant improvement in segmentation accuracy we ran the top three performing networks from our cross-validation experiment (only for the small FOV networks with the same initial whole-brain prediction) on our test data and averaged the resulting probability maps. The combined prediction of the top-three ensemble did not produce a significant improvement in accuracy and only resulted in an average of 0.3% dice improvement. Hippodeep appeared to substantially under-segment a few cases, while FIRST and FreeSurfer over-segmented multiple patients. Examples of the segmentation results for all tested methods on three participants are shown in Figure 3. The border of the hippocampus proper is depicted to highlight that the mis-segmentations by other techniques are not mainly due to differences in border definitions but to under- or over-segmentation errors of the gray matter structures. Figure 4 depicts more examples of mis-segmentations by the tested automated algorithms, specifically mis-segmentation of the hippocampus proper and segmentation of neighboring white matter structures and CS. It should be noted that our algorithm was trained on data from the same multisite studies as the first test set and that SBHV uses subjects from the SDS study as template atlases.

The cases with the highest and lowest Dice coefficients from the test dataset between manual hippocampal segmentations and our prediction (0.92 and 0.80) are presented in Figure 5. These cases highlight the quality of our output segmentations even for the case with lowest dice. The worst case demonstrates some of the challenging features for segmentation including hippocampal shrinkage, malrotation, increased ventricular volume and increased CSF surrounding the hippocampus. Additional examples of six test cases are shown in Figure S1, to demonstrate the quality of our model predictions and mismatch to manual labels. Our algorithm was two orders of magnitude faster than some of the tested techniques, segmenting the hippocampi in an average of 14 s on a GPU. While our model was on-par with the other CNN-based technique in terms of efficiency, it had a notably higher accuracy on the first test set.

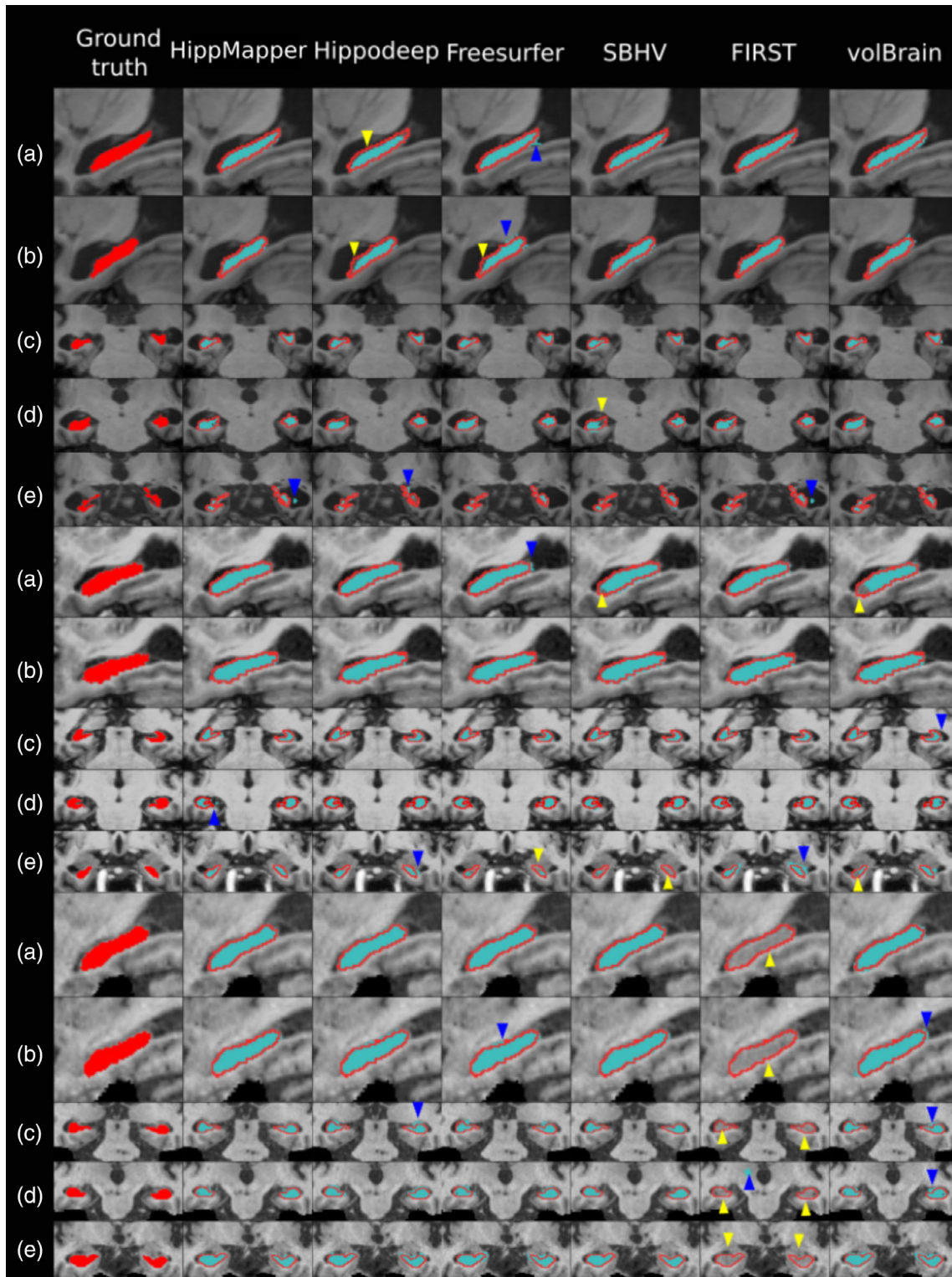
Figure 6 demonstrates cases in individuals with severe atrophy where our model produced accurate segmentation. Specifically, we

depict cases with large cysts, significant ventricular enlargement and small vessel disease. We present additional difficult cases in Figure S2, including a subject with developmental malformation and input images where the neck is occupying a large portion of the FOV. These cases are particularly problematic to atlas-based algorithms relying on registrations to a template, or those employing a registration-based initialization. While our network is able to segment both skull-stripped as well as original T1 images including the skull with comparable accuracy, for cases where the brain does not occupy a large portion of the FOV it may be optimal to skull-strip the input T1 for improved segmentations.

### 3.2 | Outlier rates

The models were also evaluated for performance, specifically outliers and failure rates, on two additional populations that are difficult to segment. Due to the lack of ground truth, the z-scores of the surface areas for each method relative to the computed subject-specific mean were used as a metric for comparison. High or low surface area z-scores (for each method) suggested that there had either been significant over or under-segmentation of the hippocampus, respectively. We considered outliers to be values above or below 2 standard deviations from the mean. Table 3 summarizes the outlier rates (average of both hemispheres) on the two populations across all the methods. volBrain had the greatest number of outliers for the FTD cohort, followed by Hippodeep (Figure 6). This population is characterized by marked hippocampal atrophy and shrinkage particularly in the left temporal lobe (where language lateralizes in most participants), which leads to common segmentation errors due to either substantial underestimation of the volume (e.g., Hippodeep's CNN-based segmentation) or overestimation with atlas-based methods (FIRST and FreeSurfer) as shown in Figure 7b. In the VCI cohort, characterized by an increased burden of WMH, vascular lesions and the presence of strokes, FreeSurfer had the greatest number of outliers and SBHV under-segmented around 10% of the cases (Figure S3). HippMapp3r had the fewest number of outliers in the two populations with a 1% outlier rate (Table 3). Example cases where our model did not produce optimal segmentations are

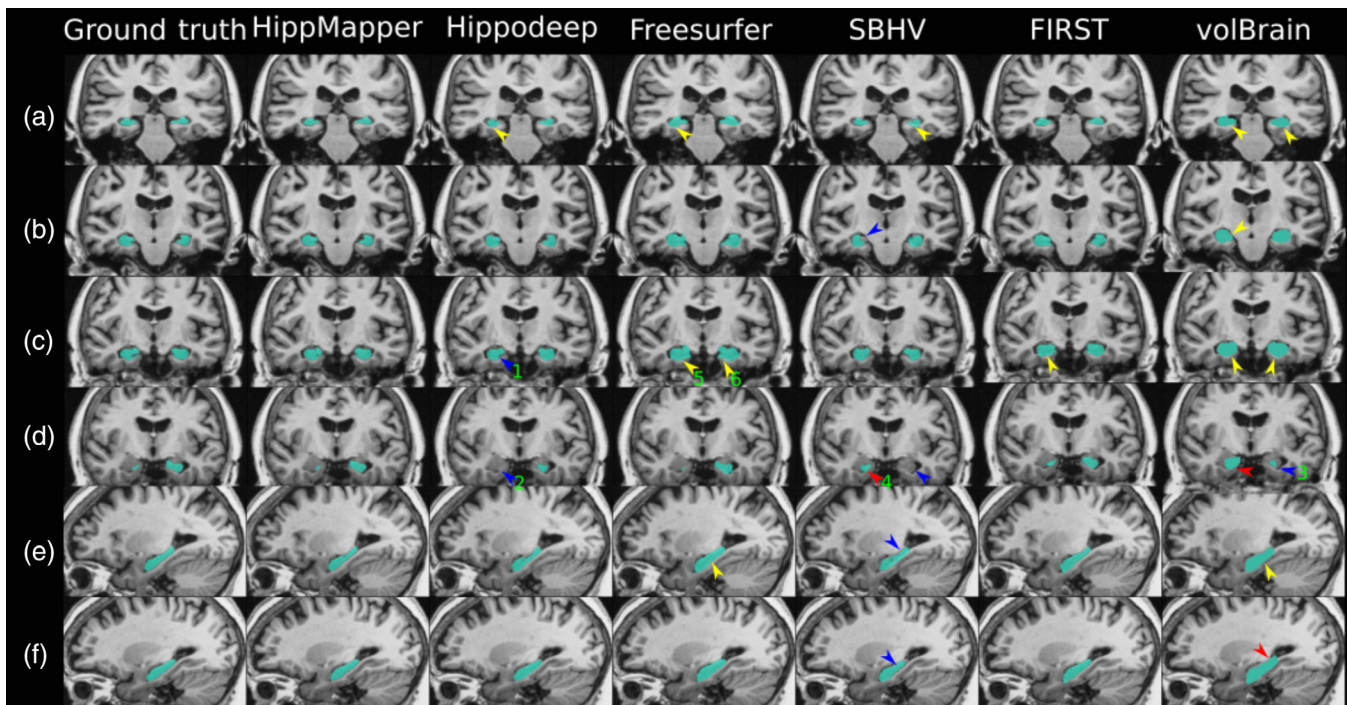




**FIGURE 3** Visual comparison of the five tested segmentation methods on three example subjects. Rows a and b display sagittal slices of the left and right hippocampi, respectively, while rows c through e display segmentations in coronal slices. Manually traced ground truth is displayed in red on the far-left column. An outline of the ground truth is overlaid on each segmentation output. Regions where over segmentation occurred are indicated by blue arrows. Regions where under segmentation occurred are indicated by yellow arrows

presented in Figure S4 and included instances with very low SNR or when a lesion localized within the hippocampus. volBrain failed on 20 VCI subjects and 11 FTD subjects. Of all completed cases on

both cohorts, volBrain had an outlier rate of 21% in the VCI cohort but failed to produce accurate results on the FTD cohort with an outlier rate close to 50%.



**FIGURE 4** Visual comparison of the six tested segmentation methods on one example subject. Rows a through d display segmentations in coronal slices, while rows e and f display sagittal slices of the left and right hippocampi, respectively. Regions where over segmentation into neighboring white matter occurred are indicated by yellow arrows. Regions where under and over segmentation of hippocampal proper occurred are indicated by blue and red arrows, respectively. Blue arrow “1” highlights missing subiculum (row c.). Blue arrow “2” indicates missing hippocampal head (row d.). Blue arrow “3” indicates under segmented hippocampal body (row d.). Red arrow “4” indicates an over segmented hippocampal head (row d.). Yellow arrows “5” and “6” show examples of over segmented white matter beyond the subiculum (row c.)

### 3.3 | Clinical adversarial attacks

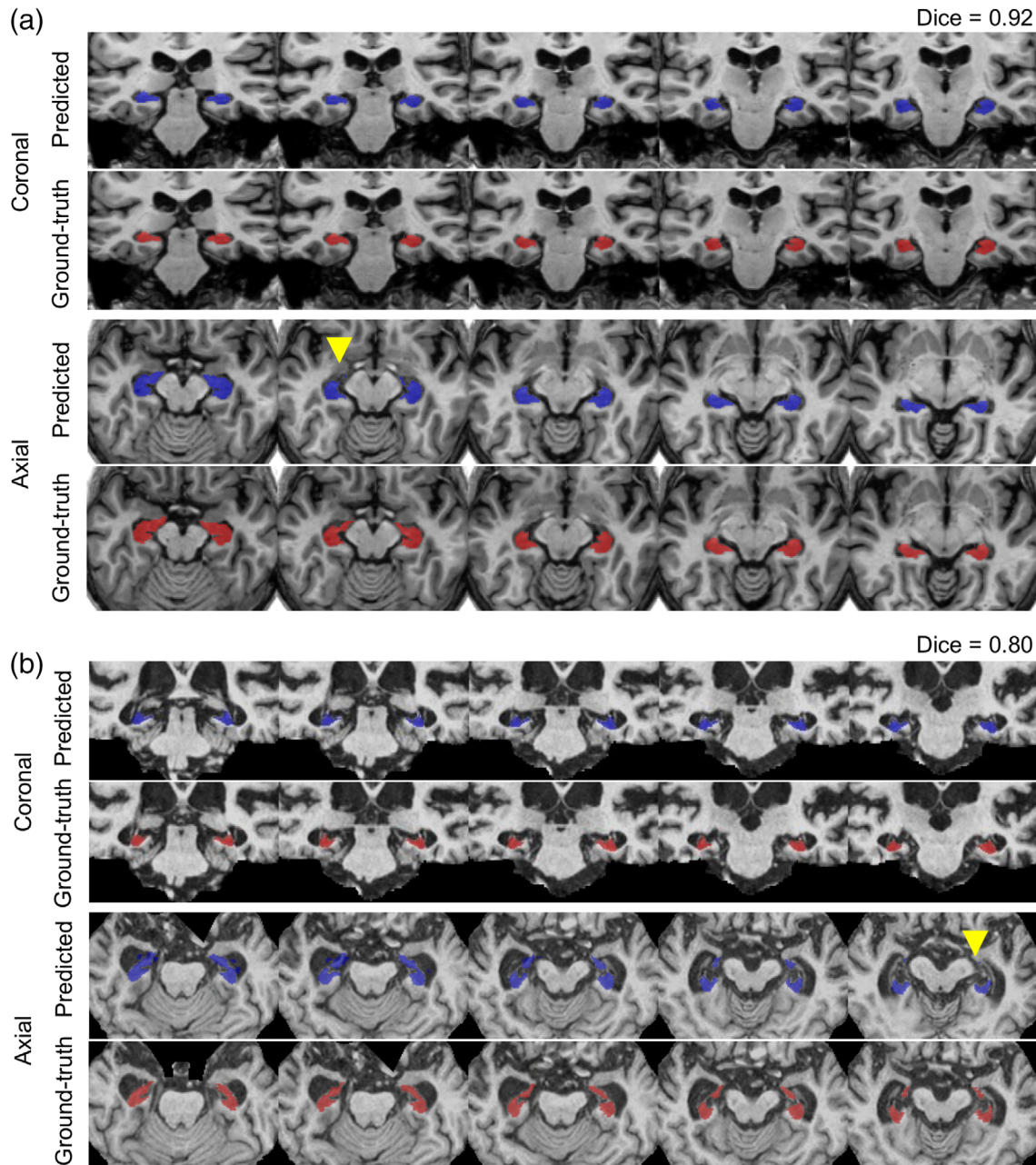
Accuracy of neural networks suffers when unseen features and characteristics from a new dataset are provided as a new test set, such as differences in image resolution, contrast, noise-level, and FOV. Clinical data are commonly obtained at lower field strengths and during shorter acquisitions than research scans, producing lower resolution and SNR data. We performed further validation experiments on both real corrupt “adversarial” cases (due to poor quality, artifacts or patient movement) and simulated adversarial data. The surface area z-scores as well as some example segmentation results for three subjects from the real adversarial cases are shown in Figure 8. None of the methods exceeded an outlier rate of 10% on these real adversarial cases (Table 3). FreeSurfer and SBHV generated the most outliers, while HippMapp3r had the lowest outlier rate (2%). We also present two corrupt scans, not included in the test datasets, where our model produced surprisingly good segmentations given that a substantial portion of the brain was not imaged at scan time (Figure S5).

The simulation experiments demonstrated that our model is robust in general to such attacks mirroring our results on the real adversarial cases, as shown in Figure 9. Our model was particularly robust to downsampling of images to 2× and cropping FOV in the Superior–Inferior plane by 30%. Downsampling to 2× in-plane and 4× out-of-plane resulted in a drop of 5% in Dice coefficient (Figure 9a).

While our model was robust to the addition of a small amount of speckle or salt-and-pepper noise (with a low standard deviation), it proved more sensitive to the addition of a large amount (sigma) of speckle noise producing a Dice coefficient drop of around 14% (Figure 9b).

## 4 | DISCUSSION

This work presents a hippocampal segmentation algorithm (HippMapp3r) based on an ensemble of 3D deep artificial neural networks. We compared our segmentation results to five state-of-the-art methods and demonstrated that the proposed algorithm is capable of producing accurate and fast hippocampal segmentations across a diverse range of neurodegenerative diseases. On a test set with manual ground truth, our algorithm achieved the highest volume correlation, Dice and Jaccard scores relative to other tested algorithms and had the highest computational speed, segmenting the hippocampus in an average of 14 s per subject. We further highlighted its robustness to atrophies and lesion types not present in the training set by demonstrating its low outlier rate on two additional test populations. We finally validated its robustness against corrupt and challenging scans on both real (with motion artifacts and low SNR) and simulated adversarial cases (through the systematic degradation of input image quality

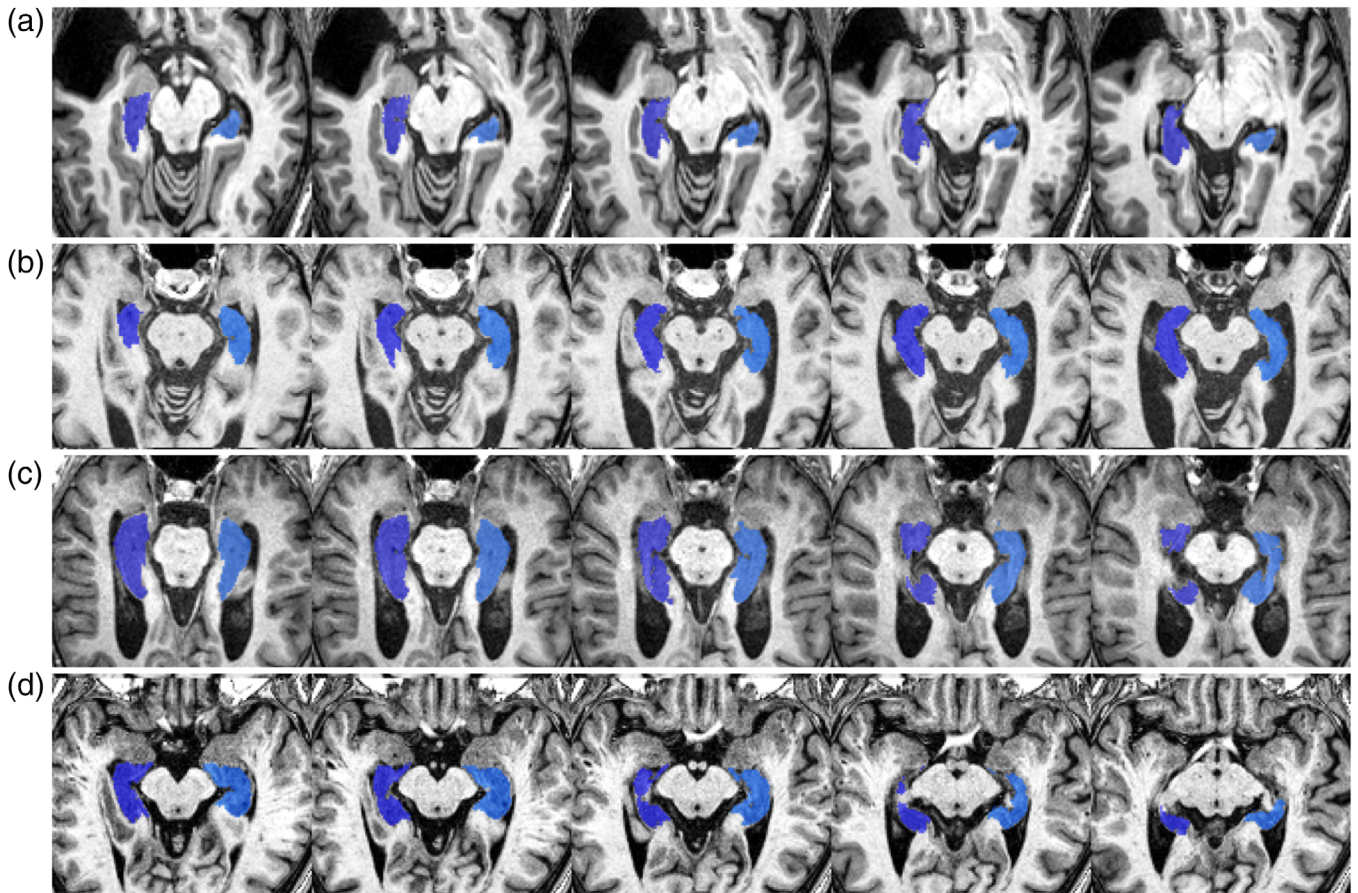


**FIGURE 5** Hippocampal segmentation cases with the highest (a) and lowest (b) dice coefficients from the test set in coronal and axial views. Blue labels represent predicted segmentations and red represent manual delineations. Yellow arrowheads highlight areas of mis-segmentations. (b) Outlines of the segmentations are presented to show the underlying image intensities and features, and demonstrate the agreement in segmentation borders. (c) Mis-segmented voxels highlight the discrepancies between the segmentations. Red voxels are manually labeled voxels not predicted by the model, and light blue voxels are predicted voxels that were not present in the manual labels

with simulation akin to low-quality clinical grade MRI). The high accuracy and efficiency of our model highlight its utility as a tool for the analysis of large multisite studies while providing the opportunity for personalized assessments.

HippMapp3r produced an average Dice value of  $0.869 \pm 0.033$  across both hippocampi ( $0.869 \pm 0.030$  left,  $0.868 \pm 0.036$  right), and Pearson's correlation coefficients of 0.95 and 0.93 for the left and right hippocampi, respectively. The high overlap and similarity

between the ground truth and the segmentation results indicate HippMapp3r's ability to segment hippocampi that are variant in shape and position across cohorts. Of note, our hippocampal model was trained on multiple segmentation protocols (in an effort to increase our training set) which have slightly different border definitions for manual segmentation than those reported within a single study. We deliberately chose to test our algorithm on images from elderly patients with brain injury, WMH, and stroke, in order to perform a



**FIGURE 6** Examples of difficult cases to segment where our model produced accurate segmentations. (a) Presence of a large cyst (top left corner) in the temporal lobe causing hippocampal rotation. (b, c) Two individuals with enlarged ventricles and hippocampal shrinkage. (d) An individual with small vessel disease and visible lacunes within the right hippocampus and surrounding white matter

**TABLE 3** Outlier rates based on surface area z-scores for all tested methods across three cohorts

Cohort	Method					
	HippMapp3r (%)	HippoDeep (%)	FIRST (%)	FreeSurfer (%)	SBHV (%)	volBrain <sup>a</sup> (%)
FTD	1	11	8	9	7	46
VCI	1	4	6	22	9	21
Real adversarial cases	2	3	4	10	10	4

Note: Rates were combined over both hemispheres.

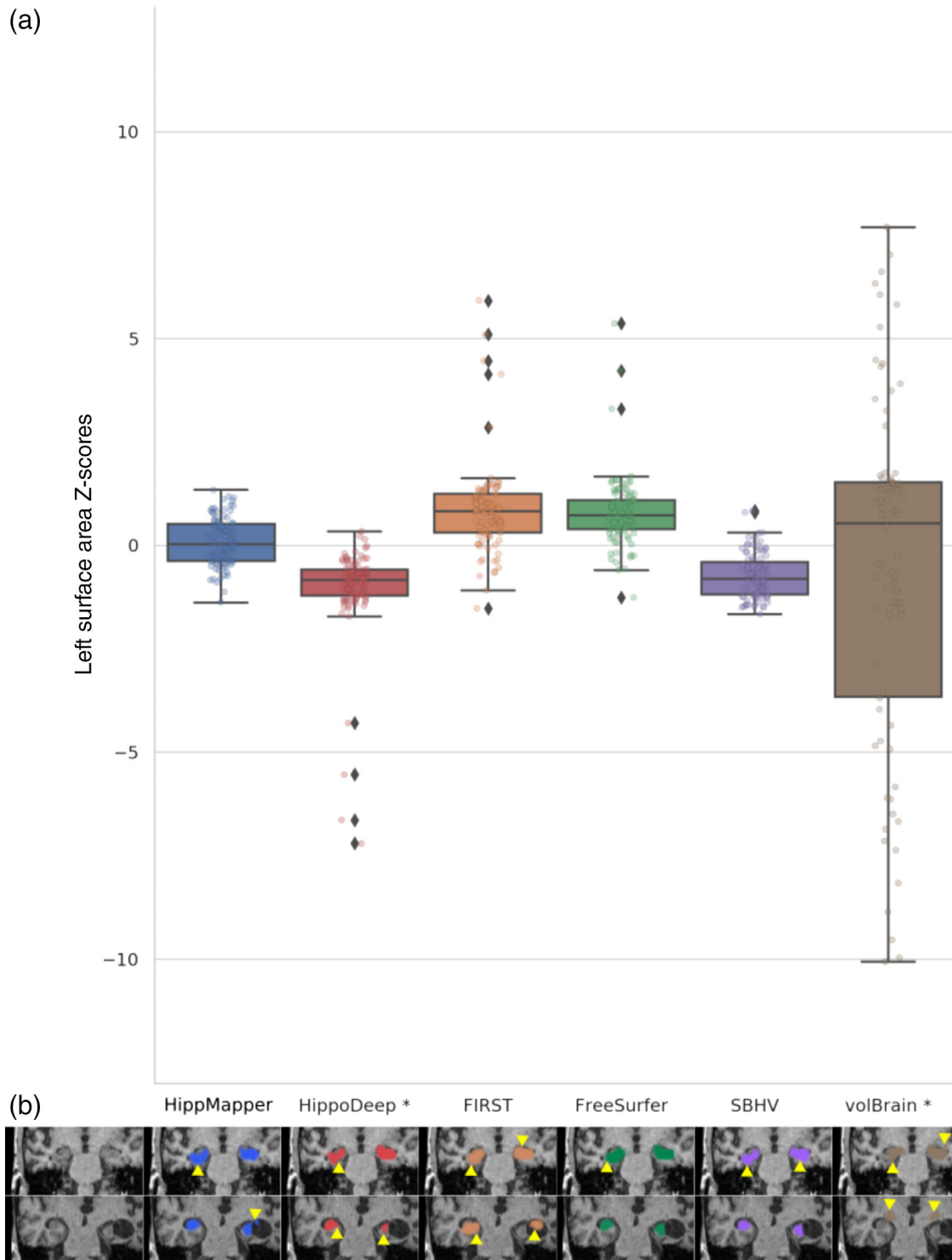
Abbreviations: FTD, frontotemporal dementia; VCI, vascular cognitive impairment.

<sup>a</sup>volBrain failed on 20 VCI subjects and 11 FTD subjects.

more clinically relevant validation (unlike many other tools reported in the literature that only use healthy young adults for validation). The presence of these cases with large atrophy, smaller volumes, hippocampal rotation and brain lesions in the test dataset may be why we obtained a lower accuracy with the state-of-the-art methods than reported elsewhere. The high variability of image quality, resolutions, acquisition parameters, and scanners in the dataset is another challenging aspect for many segmentation algorithms that were accounted for in this model by using multisite and multi-scanner studies in the training and testing datasets. Study overlap in the training and testing

set may have set our model at an advantage compared to the other methods but was unavoidable due to lack of available datasets with expert ground truth segmentations. Although training for the hippocampus model was performed using data from three different studies, it would be optimal to test it on data with manual ground truth segmentations from other studies not part of the training set.

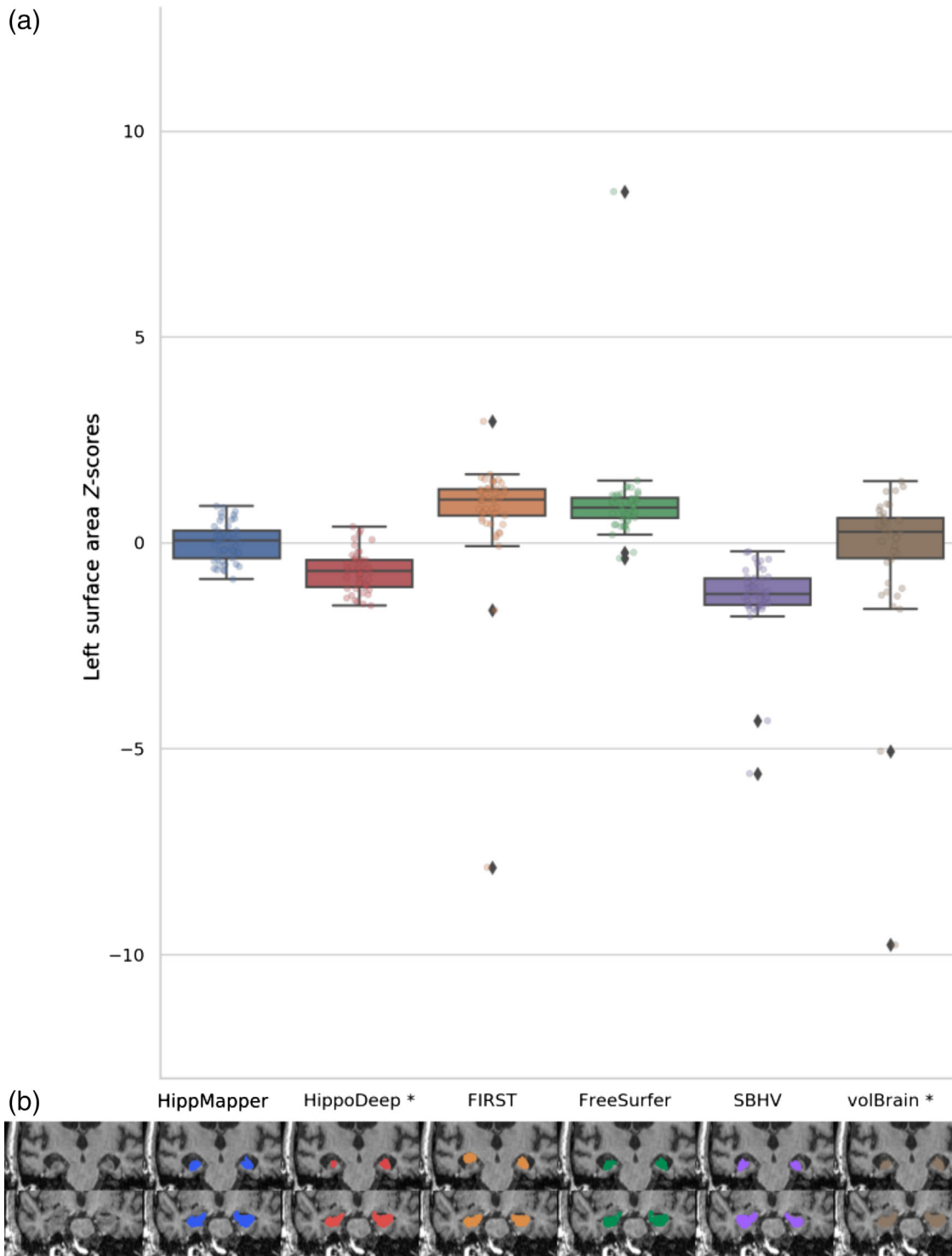
For the second test set, HippMapp3r had the fewest number of outliers, indicating it performed most consistently compared to the other methods for these populations. FreeSurfer and First tended to over segment, while SBHV tended to under segment. When observing



**FIGURE 7** Outlier rates in the frontotemporal dementia population. (a) Mean surface area z-scores for the right hippocampus segmented by HippMapper3r (blue), HippoDeep (red), FIRST (orange), FreeSurfer (green), SBHV (purple), and volBrain (brown) relative to average of segmentation volumes. (b) Visual comparison of segmentation results for two subjects in the clinical case dataset. Each row represents a distinct subject. Areas of over, under and miss-segmentation indicated by yellow arrows

performance for the different conditions, FreeSurfer performed substantially worse with the VCI cohort. This is possibly because FreeSurfer using an atlas-based method that cannot incorporate strokes or high WMH burden. Our outlier rates analyses suggest that

atlas-based methods may be susceptible to segmentation errors when tested on populations with substantial hippocampal atrophy and large hemispheric asymmetry in geometric properties, while CNN-based methods may not generalize well to test data with

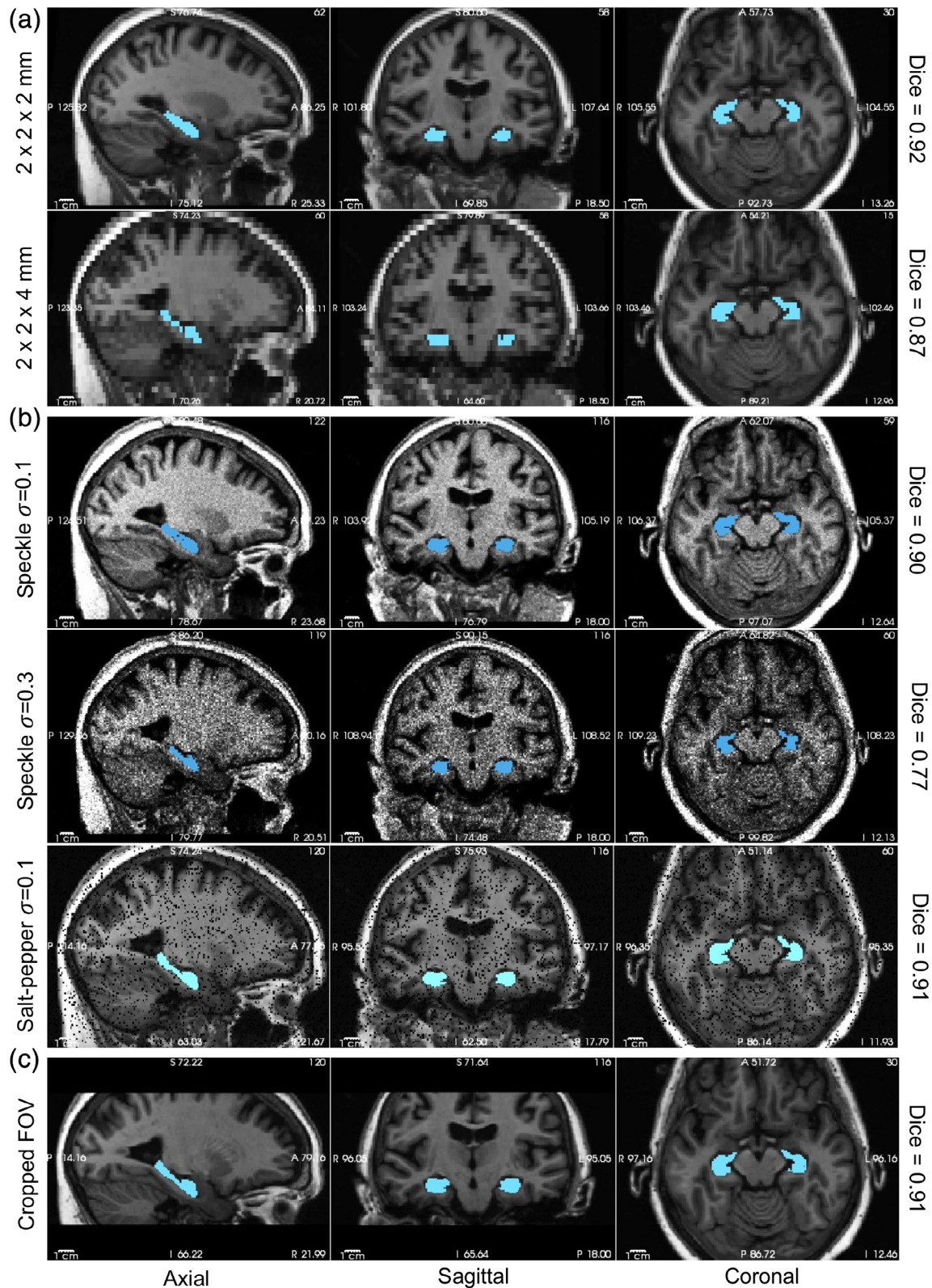


**FIGURE 8** Outlier rates on images with poor quality (real adversarial cases). (a) Mean surface area z-scores for the right hippocampus segmented by HippMapper (blue), HippoDeep (red), FIRST (orange), FreeSurfer (green), SBHV (purple), and volBrain (brown) relative to average of segmentation volumes. (b) Visual comparison of segmentation results for two subjects in the clinical case dataset. Each row represents a distinct subject. Areas of over, under and miss-segmentation indicated by yellow arrows

drastically different intensity statistics or distributions than those of the training data (for example very noisy data). While we used a filtered mean across methods to compute an estimated subject-specific mean for our outlier rates analyses, this estimate may have

been biased in the cases where most of the methods did not produce accurate segmentations.

Deep CNNs are susceptible to sharp decreases in accuracy when presented with data from different distributions than the training



**FIGURE 9** Adversarial attacks on the hippocampal model to simulate clinical data with low resolution, SNR and limited FOV, demonstrating the robustness of our model to such attacks. (a) Downsampling of resolution by  $2 \times$  in all dimensions (first row), and  $2 \times$  in-plane and  $4 \times$  out-of-plane (second row). Downsampling by  $2 \times$  did not affect segmentation accuracy, while  $2 \times$  in-plane and  $4 \times$  out-of-plane resulted in a drop of 5% in Dice coefficient. (b) Varying degrees (sigmas) of speckle and salt-and-pepper noise to simulate lower SNR. The addition of speckle noise with a large sigma produced a Dice coefficient drop of  $\sim 14\%$ . (c) Cropping of FOV in the Superior–Inferior plane by 15% in each side did not significantly affect segmentation accuracy

datasets, or with adversarial attacks. We attempted to characterize this through testing on corrupt data that failed quality control (real adversarial cases with motion and other artifacts) and simulation experiments. The low z-scores standard deviation (divergence from zero) on the real adversarial cases indicates HippMapp3r's consistency in comparison to other methods. Both the large positive deviations from the mean seen in FreeSurfer and FIRST on these cases, and the negative deviations on the part of Hippodeep, reflect the previously discussed findings when the methods were compared to ground truth labels. In the simulation experiments, we found HippMapp3r was particularly robust to the addition of speckle noise and reduced spatial resolutions. However, it was more sensitive to decreasing the FOV and exhibited a large drop in accuracy with very-noisy inputs. We did not observe any failure cases with the segmentation model on our test set or during the validation experiments.

It is worth noting that even when accurate, automated methods are used, a certain degree of neuroanatomical expertise is commonly needed to evaluate whether a segmentation is adequate or should be excluded or edited on a subject level. To enable efficient and accurate assessment of the quality of segmentations, our algorithm generates automated quality check outputs similar to those shown in Figure S6; as well as automated volumetric reports for further analyses. For group segmentations, we have added an additional function to perform outlier detection based on volumetric and shape metrics (surface area, eccentricity, and elongation), whereby subjects with metrics higher or lower than 2 standard deviations from the mean are considered outliers.

The two deep neural network algorithms (Hippodeep and HippMapp3r) were by far the more efficient with processing times on the order of seconds on a GPU as compared to several minutes or hours for the three other tools. While the FreeSurfer pipeline has a typically long computational time compared to other software, it should be noted that it also provides segmentations of numerous brain structures, as well as the hippocampal subfields. Similarly, FIRST and volBrain provided segmentations of other subcortical structures in addition to the whole hippocampus. We opted to rely on a two-network approach, with the first for initialization, as opposed to a registration-based method for initialization, to avoid incorporating registration errors and for faster predictions (even using a CPU, Video S1).

Hippodeep mainly employed hippocampus segmentations generated by FreeSurfer on a large dataset as ground truth data because manual tracings are tedious and labor-intensive (Fischl et al., 2002). Although trained using FreeSurfer segmentations, this model outperformed FreeSurfer's atlas-based algorithm by a small margin on our test dataset. Hippodeep had roughly similar computational time as HippMapp3r, but lower accuracy on the test data. This could be possibly due to the architectural differences between the two networks (including the expanding pathway, residual elements, higher number of feature maps and deep supervision in our model). It could also be due to the difference in the training data and the reliance on a bigger number of manually traced ground-truth datasets instead of a very large number of FreeSurfer segmentations.

We chose a U-net architecture as it has been shown as a successful scheme for several biomedical applications (Ronneberger et al., 2015), while also implementing a 3D design as this is advisable for volumetric medical data (Kayalibay et al., 2017; Milletari et al., 2016). In addition, we used residual units to provide smoother gradient flow through the network and skip connections to forward feature maps computed in the contracting pathway to the expanding pathway. One of the main practical limitations of applying 3D CNNs to medical data is that they are expensive to train due to their increased memory demand and thus usually require down-sampling the raw data. Our two-network "ensemble" approach avoided downsampling the hippocampal (medial temporal) region in the second pass. We opted for training our model using manually delineated ground truth segmentations by experts as they should produce more accurate results than relying on outputs from other software. Other algorithms may not fully encode the expert anatomical knowledge and may lead to biases in the predictions. We relied on augmentation through flipping and rotation but observed that augmentation through nonlinear warping did not show improvements for the training loss. We adopted the chosen parameters as they have been shown to produce optimal results in previous literature and based on prior experiments; however, further improvements in accuracy might be made if a more extensive exploration of the parameter space is performed. We ran an exploratory experiment for testing whether one could segment hippocampi and other structures straight from acquired scans, without any processing (like bias-field correction) or intensity standardization. Intensity standardization (converting image intensities to a zero-mean, 1 standard deviation) is a common preprocessing procedure in machine/deep learning known to help optimizer convergence and gradient flow. We believe that the lack of intensity standardization and the drastically different intensities across the multisite training set is the main factor for this failure in training. The results of this experiment might, however, be dependent on the training data (the histograms and intensity ranges in the set) and its size.

While the presented model is accurate and efficient, it is not flexible in terms of inputs as it requires a specific sequence (i.e., image contrasts) and thus would not produce accurate segmentations in its absence (e.g., using T2-weighted images). This is in contrast to more flexible approaches that can predict given a subset of inputs (Havaei, Guizard, Chapados, & Bengio, 2016) and thus the subject of future work. HippMapp3r also relied on the input being in a standard orientation and was unable to accurately segment when dealing with other orientations. Our segmentation algorithm does not separate the hippocampus into its different subfields, this is due to the scarcity of high-resolution sub-millimeter T1 and T2-weighted scans and manual subfield delineations in a large dataset for training, as well as lack of histological validation.

In summary, this work we present an automated whole hippocampal segmentation algorithm based on 3D CNNs that is robust to brain atrophy and lesions associated with aging and neurodegeneration of the human brain. The algorithm was trained on 209 datasets (before augmentation) from different cognitive neurology cohorts (NC, MCI, AD, or TLE), as a means of reflecting a subset of the wide range of



elderly adults that undergo brain MRI. These datasets were acquired from multiple studies to obtain good generalizability across resolutions, acquisition parameters, and scanners. Our deep networks improve upon state-of-the-art techniques in terms of both accuracy and efficiency. We observed a large margin in Dice and Jaccard similarity coefficients, and volume correlations between manual and automated segmentations for our model compared to other well-established methods. Our model was also one or two orders of magnitude faster than some of the tested methods, segmenting the hippocampi in seconds. This combination of efficiency and accuracy against other methods suggests broad utility for large multisite studies, as well as personalized assessments. We have further validated our networks by demonstrating its robustness to realistic clinic adversarial cases including sharp decreases in resolution, SNR, and cropping of FOV, indicating the potential for clinical adoption. The model is made public and accessible for use in the research setting.

## DATA ACCESSIBILITY

The developed algorithm and trained models (network weights) are publicly available at: <https://hippmapp3r.readthedocs.io>, under the GNU General Public License v3.0. An example dataset is included for testing purposes. We have developed an easy-to-use pipeline with a GUI and thorough documentation for making it accessible to users without programming knowledge.

## ORCID

Maged Goubran  <https://orcid.org/0000-0001-5880-0818>

Bradley MacIntosh  <https://orcid.org/0000-0001-7300-2355>

## REFERENCES

Anon. (2005). The human hippocampus, H.M. Duvernoy Third edition, Springer-Verlag Berlin Heidelberg 2005 (232 pages). *Journal of Neuroradiology*, 32(4), 288. [https://doi.org/10.1016/s0150-9861\(05\)83157-8](https://doi.org/10.1016/s0150-9861(05)83157-8)

Barnes, J., Bartlett, J. W., van de Pol, L. A., Loy, C. T., Scahill, R. I., Frost, C., ... Fox, N. C. (2009). A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. *Neurobiology of Aging*, 30, 1711–1723.

Bernasconi, N., Natsume, J., & Bernasconi, A. (2005). Progression in temporal lobe epilepsy: Differential atrophy in mesial temporal structures. *Neurology*, 65, 223–228.

Boccardi, M., Bocchetta, M., Morency, F. C., Collins, D. L., Nishikawa, M., Ganzola, R., ... EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative. (2015). Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimers Dement*, 11, 175–183.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science* (pp. 424–432).

Courchesne, E., Chisum, H. J., Townsend, J., Cowles, A., Covington, J., Egaas, B., ... Press, G. A. (2000). Normal brain development and aging: Quantitative analysis at in vivo MR imaging in healthy volunteers. *Radiology*, 216, 672–682.

Das, S. R., Mechanic-Hamilton, D., Korczykowski, M., Pluta, J., Glynn, S., Avants, B. B., ... Yushkevich, P. A. (2009). Structure specific analysis of the hippocampus in temporal lobe epilepsy. *Hippocampus*, 19, 517–525.

Deshpande, N. A., Gao, F. Q., Bakshi, S. N., Leibovitch, F. S., Black, S. E., & Sunnybrook Dementia Study. (2004). Simple linear and area MR measurements can help distinguish between Alzheimer's disease, frontotemporal dementia, and normal aging: The Sunnybrook Dementia Study. *Brain and Cognition*, 54, 165–166.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26, 297–302.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., ... Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33, 341–355.

Goubran, M., Bernhardt, B. C., Cantor-Rivera, D., Lau, J. C., Bliston, C., Hammond, R. R., ... Khan, A. R. (2016). In vivo MRI signatures of hippocampal subfield pathology in intractable epilepsy. *Human Brain Mapping*, 37, 1103–1119.

Goubran, M., Rudko, D. A., Santyr, B., Gati, J., Szekeres, T., Peters, T. M., & Khan, A. R. (2014). In vivo normative atlas of the hippocampal subfields using multi-echo susceptibility imaging at 7 Tesla. *Human Brain Mapping*, 35, 3588–3601.

Hasboun, D., Chantôme, M., Zouaoui, A., Sahel, M., Deladoeuille, M., Sourour, N., ... Dormont, D. (1996). MR determination of hippocampal volume: Comparison of three methods. *AJNR. American Journal of Neuroradiology*, 17, 1091–1098.

Havaei, M., Guizard, N., Chapados, N., & Bengio Y (2016). HeMIS: Heteromodal image segmentation. arXiv [cs.CV]. Retrieved from <http://arxiv.org/abs/1607.05194>

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://doi.org/10.1109/cvpr.2016.90>

Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., ... Alzheimer's Disease Neuroimaging Initiative. (2015). A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. *NeuroImage*, 115, 117–137.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., & Maier-Hein, K. H. (2018). Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge. In *Lecture notes in computer science* (pp. 287–297). Berlin, Germany: Springer Science+Business Media.

Jaccard, P. (1912). The distribution of the FLORA in the alpine ZONE.1. *The New Phytologist*, 11, 37–50.

Jack, C. R., Jr., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., ... Kokmen, E. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology*, 51, 993–999.

Jack, C. R., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., ... Kokmen, E. (2000). Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. *Neurology*, 55, 484–490.

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., ... Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36, 61–78.

Kayalibay, B., Jensen, G., & van der Smagt, P. (2017). CNN-based segmentation of medical imaging data. arXiv [cs.CV]. Retrieved from <http://arxiv.org/abs/1701.03056>

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv [cs.LG]. Retrieved from <http://arxiv.org/abs/1412.6980>

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. USA: Curran Associates Inc. NIPS'12 (pp. 1097–1105).

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2014). Deeply-supervised nets. arXiv [stat.ML]. Retrieved from <http://arxiv.org/abs/1409.5185>

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision*

- and Pattern Recognition (CVPR). <https://doi.org/10.1109/cvpr.2015.7298965>
- Manjón, J. V., & Coupé, P. (2016). volBrain: An online MRI brain volumetry system. *Frontiers in Neuroinformatics*, 10, 30.
- Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV). <https://doi.org/10.1109/3dv.2016.79>
- Nestor, S. M., Gibson, E., Gao, F.-Q., Kiss, A., Black, S. E., & Alzheimer's Disease Neuroimaging Initiative. (2013). A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in Alzheimer's disease. *NeuroImage*, 66, 50–70.
- Patenaude, B., Smith, S. M., Kennedy, D. N., & Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56, 907–922.
- Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.
- Pluta, J., Avants, B. B., Glynn, S., Awate, S., Gee, J. C., & Detre, J. A. (2009). Appearance and incomplete label matching for diffeomorphic template based hippocampus segmentation. *Hippocampus*, 19, 565–571.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science* (pp. 234–241). Berlin, Germany: Springer Science +Business Media.
- Rusinek, H., De Santi, S., Frid, D., Tsui, W.-H., Tarshish, C. Y., Convit, A., & de Leon, M. J. (2003). Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal MR imaging study of normal aging. *Radiology*, 229, 691–696.
- Santyr, B. G., Goubran, M., Lau, J. C., Kwan, B. Y. M., Salehi, F., Lee, D. H., ... Khan, A. R. (2017). Investigation of hippocampal substructures in focal temporal lobe epilepsy with and without hippocampal sclerosis at 7T. *Journal of Magnetic Resonance Imaging*, 45, 1359–1370.
- Scahill, R. I., Frost, C., Jenkins, R., Whitwell, J. L., Rossor, M. N., & Fox, N. C. (2003). A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. *Archives of Neurology*, 60, 989–994.
- Schoemaker, D., Buss, C., Head, K., Sandman, C. A., Davis, E. P., Chakravarty, M. M., ... Pruessner, J. C. (2016). Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *NeuroImage*, 129, 1–14.
- Steen, R. G., Mull, C., McClure, R., Hamer, R. M., & Lieberman, J. A. (2006). Brain volume in first-episode schizophrenia: Systematic review and meta-analysis of magnetic resonance imaging studies. *The British Journal of Psychiatry*, 188, 510–518.
- Sullivan, E. V. (2002). Differential rates of regional brain change in callosal and ventricular size: A 4-year longitudinal MRI study of elderly men. *Cerebral Cortex*, 12, 438–445.
- Thyreau, B., Sato, K., Fukuda, H., & Taki, Y. (2018). Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. *Medical Image Analysis*, 43, 214–228.
- Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29, 1310–1320.
- Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv [cs.CV]. Retrieved from <http://arxiv.org/abs/1607.08022>
- Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31, 1116–1128.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Goubran M, Ntiri EE, Akhavein H, et al. Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks. *Hum Brain Mapp*. 2020;41:291–308. <https://doi.org/10.1002/hbm.24811>