UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Engenharia Elétrica e de Computação

Oeslle Alexandre Soares de Lucena

Deep Learning for Brain Analysis in MR Imaging

Aprendizado Profundo para Análise do Cérebro em Imagens de Ressonância Magnética

Campinas

2018

Oeslle Alexandre Soares de Lucena

Deep Learning for Brain Analysis in MR Imaging

Aprendizado Profundo para Análise do Cérebro em Imagens de Ressonância Magnética

A Dissertation presented to the School of Electrical and Computer Engineering of the State University of Campinas in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering, in the field of Computer Engineering.

Dissertação de mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Supervisor: Prof. Dr. Roberto de Alencar Lotufo

Co-supervisor Profa. Dra. Letícia Rittner

Este exemplar corresponde à versão final da tese defendida pelo aluno Oeslle Alexandre Soares de Lucena, e orientada pelo Prof. Dr. Roberto de Alencar Lotufo

Campinas

2018

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2016/18332-8; FAPESP, 2017/23747-5
ORCID: https://orcid.org/0000-0002-0658-6589

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

Lucena, Oeslle Alexandre Soares de, 1992-
L963d Deep learning for brain analysis in MR imaging / Oeslle Alexandre Soares de Lucena. – Campinas, SP : [s.n.], 2018.

Orientadores: Roberto de Alencar Lotufo e Sebastien Ourselin.
Coorientador: Leticia Rittner.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.
Em regime interinstitucional com: University College London.

1. Ressonância magnética. 2. Cérebro. I. Lotufo, Roberto de Alencar, 1955-. II. Ourselin, Sebastien. III. Rittner, Leticia, 1972-. IV. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. VI. Título.

Informações para Biblioteca Digital

Título em outro idioma: Aprendizado profundo para análise do cérebro em imagens de ressonância magnética
Palavras-chave em inglês:
Magnetic resonance
Brain
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Roberto de Alencar Lotufo [Orientador]
Nina Sumiko Tomita Hirata
José Mario de Martino
Data de defesa: 11-09-2018
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidato: Oeslle Alexandre Soares de Lucena **RA:** 190715
Data da defesa: 11/09/2018

Dissertation Title: "Deep Learning for Brain Analysis in MR Imaging".

Titulo da Dissertação: "Aprendizado Profundo para Análise do Cérebro em Imagens de Ressonância Magnética".

Prof. Dr. Roberto de Alencar Lotufo (Presidente, FEEC/UNICAMP)
Profa. Dra. Nina Sumiko Tomita Hirata (IME/USP)
Prof. Dr. José Mario de Martino (FEEC/UNICAMP)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no processo de vida acadêmica do aluno.

Acknowledgements

Firstly, I would like to thank my family: my dad Oscar, my mom Izabel, my sister Izabelle, my brother from another mother Robson, and my other brother from another mother Wesley for absolutely every single support in my life. You all have a special place in my heart, and with no doubt are responsible for pushing me forward to achieve my dreams!

Secondly, I would like to express my sincere gratitude to Roberto Souza for advising me during my M.Sc. study, believe in myself when I could not do it, teach how to be a good researcher and a respectful human being. Thank you very much, my friend! I am really glad we crossed paths. You are my role model!

Secondly, I would like to thank to my supervisor Prof. Lotufo and my co-supervisor Dr. Rittner for the continuous support of my M.Sc. study and related research and immense knowledge. Your guidance helped me in all the time of research and writing of this dissertation. In particular, I want to reinforce my appreciation to Dr. Lotufo advising.Thank you very much for giving the opportunity for me to come here and work under your supervision. I feel honored and I could not have imagined having a better advisor.

I thank my fellow LCA/MICLab labmates in for the stimulating discussions and for all the fun we have had in the last two years. Also I thank my friends from UFCG: Luciana Joviniano, Geraldo Landim, Lucas Moreira, Lucas Henrique, Felipe Pontes, Erico Castro, and Clarisse Petua. Without you guys, I would never have gone this far. In particular, I am grateful to Dr. Waslon Terlizzie for enlightening me the first glance of research.

Last but not the least, my sincere thanks also goes to Prof.Sebastien Ourselin, Sjoerd Vos, Rachel Sparks who provided me an opportunity to join their team as intern at UCL, and who gave access to the laboratory and research facilities. Without they precious support it would not be possible to conduct this research. Also, I thank my UCL colleagues: Alejandro, Ioana, Maria, Daniele, and Michael, and Charles Bell's House Admin team: Kate, Katie, and Sarah for all the support, love and kindness that I received during my internship.

"For me, it is far better to grasp the Universe as it really is than to persist in delusion, however satisfying and reassuring."

(Carl Sagan)

Abstract

Convolutional neural networks (CNNs) are one branch of deep learning that have performed successfully in many brain magnetic resonance (MR) imaging analysis. CNNs are representation-learning methods with stacked layers comprised of a convolution operation followed by a non-linear activation and pooling layers. In these networks, each layer outputs a higher and more abstract representation from a given input, in which the weights of the convolutional layers are learned by an optimization problem. In this work, we tackled two problems using deep-learning-based approaches: skull-stripping (SS) and tractography. We firstly proposed a full CNN-based SS trained with what we refer to as silver standard masks. Segmenting brain tissue from non-brain tissue is a process known as brain extraction or skull-stripping. Silver standard masks are generated by forming the consensus from a set of eight, public, non-deep-learning-based SS methods using the algorithm Simultaneous Truth and Performance Level Estimation (STAPLE). Our approach reached state-of-the-art performance, generalized optimally, decreased inter-/intra-rater variability, and avoided CNN segmentation overfitting towards one specific manual annotation. Secondly, we investigated a CNN-based tractography solution for epilepsy surgery. The main goal of this analysis was to structure a baseline for a deep-learning-based-regression to predict white matter fiber orientations. Tractography is a visualization of the white matter fibers or tracts; its goal in presurgical planing is simply to identify the position of eloquent pathways, such as the motor, sensory, and language tracts to reduce the risk of damaging these critical structures. We performed analysis cross-validation using only in a single patient per time, and also, training with data from 10 patients for training the CNN. Our results were not optimal, however, the tracts tended to be of a similar length and converged to the mean fiber tract locations. Additionally, to the best of our knowledge, our method is the first approach that investigates CNNs for tractography, and thus, our work is a baseline for this topic.

Keywords: Deep Learning; Convolutional Neural Network; Skull-stripping; Tractography; Magnetic Resonance Image; Silver-standard Masks.

Resumo

Redes neurais convolucionais (CNNs-*Convolutional neural networks*) são uma vertente do apredizado profundo que obtiveram muito sucesso quando aplicadas em várias análises em imagens de ressonância magnética (MR-*magnetic resonance*) do cérebro. As CNNs são métodos de aprendizagem de representação com várias camadas empilhadas compostas por uma operação de convolução seguida de uma ativação não linear e de camadas de agrupamento. Nessas redes, cada camada gera uma representação mais alta e mais abstrata de uma determinada entrada, na qual os pesos das camadas convolucionais são aprendidos por um problema de otimização. Neste trabalho, tratamos dois problemas usando abordagens baseadas em aprendizagem profunda: remoção da calota craniana (SS) e tractografia. Primeiramente, propusemos um SS completo baseado em CNN treinado com o que nos referimos como máscaras de padrão de prata. A segmentação de tecido cerebral a partir de tecido não cerebral é um processo conhecido como extração da calota craniana ou remoção de crânios. As máscaras de padrão de prata são geradas pela formação do consenso a partir de um conjunto de oito métodos de SS públicos, não baseados em aprendizagem profunda, usando o algoritmo Verdade Simultânea e Estimativa do Nível de Desempenho (STAPLE-*Simultaneous Truth and Performance Level Estimation*). Nossa abordagem alcançou o desempenho do estado da arte, generalizou de forma otimizada, diminuiu a variabilidade inter / intra-avaliador e evitou a super-especialização da segmentação da CNN em relação a uma anotação manual específica. Em segundo lugar, investigamos uma solução de tractografia baseada em CNN para cirurgia de epilepsia. O principal objetivo desta análise foi estruturar uma linha de base para uma regressão baseada em aprendizagem profunda para prever as orientações da fibra da matéria branca. Tractografia é uma visualização das fibras ou tratos da substância branca; seu objetivo no planejamento pré-operatório é simplesmente identificar a posição de caminhos eloqüentes, como os tratos motor, sensorial e de linguagem, para reduzir o risco de danificar essas estruturas críticas. Realizamos uma análise em um único paciente e também uma análise entre 10 pacientes em uma abordagem de validação cruzada. Nossos resultados não foram ótimos, entretanto, as fibras preditas pelo algoritmo tenderam a ter um comprimento similar e convergiram para os locais médios do trato das fibras. Além disso, até onde sabemos, nosso método é a primeira abordagem que investiga CNNs para tractografia, e assim, nosso trabalho é uma base para este tópico.

Palavras-chaves: Aprendizado Profundo; Redes Neurais Convolucionais; Extração de Crânio; Tractografia; Imagem de Resonância Magnética; Máscara de Padrão Prata.

List of Figures

Figure 1 – Representative 3D reconstruction of the manual (gold standard) annotation for one subject of the *CC-359*, OASIS and LPBA40 datasets. . . 23

Figure 2 – Modified RECOD U-Net architecture. (a) presents configurations of the contractive path blocks (CPB), connection blocks (CB), and expansive path blocks (EPB) used in the architecture. (b) illustrates the entire architecture consisting of CPB, CB, and EPB modules. The contracting path is on the left and the expansive path is on the right. The concatenations (red arrows) are always done between the output of the third convolutional layer (red in CPB block of (a)) in the contracting path and the output of the previous block in the expansive path. The number of filters in two first EPB blocks vary for each convolution. The text in blue and orange of (b) correspond to the convolution layers blocks of the same colors in the EPB block in (a). 27

Figure 3 – Proposed deep learning brain segmentation pipelines. Both pipelines consist of three stages: pre-processing (purple), CNN segmentation (green), and threshold/post-processing (red). The CONSNet pipeline is shown in (a), and the auto-context CONSNet pipeline is shown in (b). The blue box in Figure (b) represents the probability generation done by CONSNet to be used as input in auto-context CONSNet. 29

Figure 4 – Representative 3D reconstruction of the different segmentation methods for one subject of the *CC-12* subset. 33

Figure 5 – Representative 3D reconstruction of the different segmentation methods for one subject of the LPBA40 dataset. 34

Figure 6 – Representative 3D reconstruction of the different segmentation methods for one subject of the OASIS dataset. 35

Figure 7 – Box plots of average Dice coefficient for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data. . . . 37

Figure 8 – Box plots of average sensitivity for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data. 37

Figure 9 – Box plots of average specificity for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data. 38
Figure 10 – Box plots of average Hausdorff distance for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data. . . . 38
Figure 11 – Box plots of average symmetric surface-to-surface mean distance for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data. 39
Figure 12 – Heat-map of the p-values calculated for the auto-context CONSNet across all evaluation metrics assessed in the *CC-12* subset. Darker cells highlight statistical significance (p-values < 0.05). 40
Figure 13 – Heat-map of the p-values calculated for the auto-context CONSNet across all evaluation metrics assessed in the LPBA40 dataset. Darker cells highlight statistical significance (p-values < 0.05) 41
Figure 14 – Heat-map of the p-values calculated for the auto-context CONSNet across all evaluation metrics assessed in the OASIS dataset. Darker cells highlight statistical significance (p-values < 0.05) 41
Figure 15 – Sagittal, coronal and axial heat map projections for the *CC-12* subset showing (a) false positive (FP) and (b) false negative (FN). The manual segmentations was used as the reference. Brighter voxels represents a high systematic number of FPs or FNs. 42
Figure 16 – Sagittal, coronal and axial heat map projections for the LPBA40 dataset showing (a) false positive (FP) and (b) false negative (FN). The manual segmentations was used as the reference. Brighter voxels represents a high systematic number of FPs or FNs. 43
Figure 17 – Sagittal, coronal and axial heat map projections for the OASIS dataset showing (a) false positive (FP) and (b) false negative (FN). The manual segmentations was used as the reference. Brighter voxels represents a high systematic number of FPs or FNs. 44
Figure 18 – 3D reconstruction of the UF tract. 53
Figure 19 – 18-layer 3D ResNet implemented. (a) Presents the residual blocks implemented in the network. (b) Depicts the full architecture adopted where FC stands for fully connected layer. 55

Figure 20 – Proposed tractography pipeline. (a) Autoencoder , (b) regression training to map the high-level embeddings representation, and (c) fine-tuning the whole network (ResNet18 + trained decoding layers) 56
Figure 21 – Bar plot showing the overlap differences computed by the Dice in Experiment 1 and 2. The vertical lines in black are a measure of the standard deviation per subject along the experiments. 59
Figure 22 – Representative 3D reconstruction of the left UF tract in Experiment 1, which the ground truth annotation is in red and the prediction from our method is yellow. (a)-(e) 3D reconstructions of the subject **6** (best performance among all subjects) and (f)-(j) subject **9** (worst performance among all subjects) 60
Figure 23 – Representative 3D reconstruction of the left UF tract in Experiment 2, which the ground truth annotation is in red and the prediction from our method is yellow. (a)-(e) 3D reconstructions of the subject **5** (best performance among all subjects) and (f)-(j) subject **7** (worst performance among all subjects) 61

List of Tables

Table 1 – Overall analysis against manual segmentation results for the *CC-12* subset. The best two values and all values better than auto-context CONSNet are emboldened. SSSMD = symmetric surface-to-surface mean distance. 34
Table 2 – Overall analysis against manual segmentation results for the LPBA40 dataset. The best two values and all values better than auto-context CONSNet are emboldened. SSSMD = symmetric surface-to-surface mean distance. 35
Table 3 – Overall analysis against manual segmentation results for the *OASIS* dataset. The best two values and all values better than auto-context CONSNet are emboldened. SSSMD = symmetric surface-to-surface mean distance. 36
Table 4 – Processing times for one image volume of each dataset (*CC-359*, OASIS, and LPAB40) for each skull-stripping method. For the CONSNet approaches, the number in front of the backslash represents the time computed on the CPU while the number after the backslash is the GPU time. * denotes results for the processing time for the STAPLE consensus-forming step only or the auto-context CONSNet step only (see text). 45
Table 5 – Two-fold cross-validation using the LPBA40 dataset. The best score using the LPBA40 dataset for each metric is emboldened. Values for 3D CNN (KLEESIEK *et al.*, 2016), and for auto-net and U-Net (SALEHI *et al.*, 2017) are from literature. SSSMD = symmetric surface-to-surface mean distance. 46
Table 6 – Two-fold cross-validation using the OASIS dataset. The best score using the OASIS dataset for each metric is emboldened. Values for 3D CNN (KLEESIEK *et al.*, 2016), and for auto-net and U-Net (SALEHI *et al.*, 2017) are from literature. SSSMD = symmetric surface-to-surface mean distance. 46
Table 7 – Averaged results for five-fold cross-validation in experiments 1 and 2. The best scores are emboldened and the worst are underlined. 58

List of Acronyms

BEaST	Brain Extraction based on Non-local Segmentation Technique
BET	Brain Extraction Tool
BSE	Brain Surface Extractor
CC-359	Calgary-Campinas-359
CNNs	Convolutional Neural Networks
CRFs	Conditional Random Fields
CSD	Constrained Spherical Deconvolution
DL	Deep Learning
dMRI	Diffusion Magnetic Resonance Image
EZ	Epileptogenic Zone
FCNs	Fully Convolutional Networks
GE	General Electric
HWA	Hybrid Watershed Approach
IBSR	Internet Brain Segmentation Repository
LPBA40	LONI Probabilistic Brain Atlas
MBWSS	Marker Based Watershed Scalper
MR	Magnetic Resonance
OASIS	Open Access Series of Imaging Studies
OPTIBET	Optimized Brain Extraction
ROBEX	Robust Brain Extraction
STAPLE	Simultaneous Truth and Performance Level Estimation
SS	Skull-stripping
WM	White Matter
UF	Uncinate Fasciculus

Contents

1 Introduction 16
1.1 Motivation 16
1.2 Objectives 17
1.3 Contributions 17
1.3.1 Skull-stripping 17
1.3.2 Tractography 18
1.4 Organization of the Dissertation 18
2 Convolutional Neural Networks for Skull-stripping in Brain MR Imaging using Consensus-based Silver standard Masks 19
2.1 Motivation 19
2.1.1 Previous Works in Skull Stripping 20
2.1.2 Our Approach 22
2.2 Materials and Methods 23
2.2.1 Datasets 23
2.2.1.1 *CC-359* Dataset 23
2.2.1.2 LPBA40 Dataset 24
2.2.1.3 OASIS Dataset 24
2.2.2 Automatic Skull-stripping Methods 24
2.2.3 STAPLE-derived Silver Standard Consensus 25
2.2.4 Convolutional Neural Network Architecture and Implementation . . 26
2.2.5 Proposed Brain Extraction Pipelines 26
2.2.5.1 Pre-processing Step 28
2.2.5.2 CNN Segmentation Step 28
2.2.5.3 Threshold and Post-processing Step 30
2.2.6 Evaluation Metrics and Statistical Analysis 30
2.2.7 Experimental Methodology 31
2.2.7.1 Experiment 1 32
2.2.7.2 Experiment 2 32
2.3 Results 32
2.3.1 Experiment 1 32
2.3.2 Experiment 2 45
2.4 Discussion 45
2.4.1 Experiment 1 45
2.4.2 Experiment 2 48
2.4.3 General Discussion 48

2.5 Conclusions . . . 49
3 Deep-learning-based Tractography for Surgical Planning in Epilepsy treatment 50
3.1 Motivation . . . 50
3.1.1 Previous Works in Tractography . . . 50
3.1.2 Our Approach . . . 52
3.2 Materials and Methods . . . 52
3.2.1 Dataset . . . 52
3.2.2 Convolutional Neural Network Architectures and Implementation . 53
3.2.3 Proposed Tractography Pipeline . . . 54
3.2.3.1 Pre-processing . . . 54
3.2.3.2 CNN Regression . . . 56
3.2.4 Experimental Methodology . . . 57
3.2.5 Evaluation Metrics . . . 57
3.3 Results . . . 57
3.4 Discussion . . . 58
3.4.1 Experiment 1 . . . 58
3.4.2 Experiment 2 . . . 59
3.4.3 General Discussion . . . 59
3.5 Conclusions . . . 60
4 Conclusions . . . 62
Conclusions . . . 62
4.1 Publications . . . 63
4.1.1 Journal Papers . . . 63
4.1.2 Conference Papers . . . 63
4.1.3 Abstracts . . . 64

Bibliography . . . 65

1 Introduction

1.1 Motivation

In medical image analysis, magnetic resonance (MR) imaging modality is widely used in clinical medicine and medical research, especially in diagnosing and studying brain disorders. MR imaging exhibits excellent soft tissue contrast that is not usually found in other imaging modalities, such as x-ray or computed tomography. As a consequence, MR scanning is broadly accepted as producing good-to-excellent visualization of brain structures as an example (BROWN *et al.*, 2014). Moreover, MR is often used to diagnose a variety of brain diseases including acoustic neuroma, Alzheimer's disease and other neurodegernative conditions, cerebrovascular diseases (like brain aneurysm, arteriovenous malformations, stroke), and tumours (KALAVATHI; PRASATH, 2016).

In this field, deep-learning-based approaches, in particular convolutional neural networks (CNNs) have been widely used in a myriad of tasks (GREENSPAN *et al.*, 2016; LITJENS *et al.*, 2017; KER *et al.*, 2018; RAVÌ *et al.*, 2017), such as: brain structures segmentation (BREBISSON; MONTANA, 2015), lesion detection (HUANG *et al.*, 2017), classification of abnormalities (BAR *et al.*, 2015), reconstruction and registration (CHENG *et al.*, 2018; HAMMERNIK *et al.*, 2018). CNNs are representation-learning methods with multiple stack layers comprised of a convolution operation followed by a non-linear activation and pooling layers (LECUN *et al.*, 2015; LITJENS *et al.*, 2017). In these networks, each layer outputs a higher and more abstract representation from a given input (SCHMIDHUBER, 2015), in which the weights of the convolutional layers are learned by an optimization problem. Therefore, opposed to conventional machine learning algorithms, CNNs can discover automatically the representations needed for a specific task by feeding raw data as input (i.e. classification, segmentation, regression, etc), without the need of handcrafting features (LECUN *et al.*, 2015).

Although CNNs have performed successfully in brain medical image analysis, training a CNN from scratch in a end-to-end solution is still a challenge due to the amount of annotated data needed for the supervised training (GREENSPAN *et al.*, 2016; LUCENA *et al.*, 2018). The "gold-standard" annotation in medical image datasets is usually done by a manual rater and this stage can be time-consuming and expensive. Also, manual annotation is a subjective procedure. For instance, there are more than a dozen protocols for hippocampus segmentation and different protocols provide up to 2.5-fold volume difference (SOUZA *et al.*, 2018). As a consequence, intra/inter rater variability affect the validation results in such data driven methods. In order to mitigate inter-/intra-rater variability and a potential alternative for the lack of "gold-standard" labeled data,

the consensus of multiple automatic methods could be used in CNN-based methods.

CNNs have been also employed to in surgical planning achieving optimal results (HUFF *et al.*, 2018; IBRAGIMOV *et al.*, 2017; WANG *et al.*, 2018). A particular case is a pre-surgical planning of the epileptogenic zone surgery, which is the minimal area of cortex that must be resected to produce seizure-freedom (LÜDERS *et al.*, 2006). For this case, tractography is used to identify the position of eloquent pathways, such as the motor, sensory, and language tracts (BERMAN, 2009) to plan the surgery aiming to avoid damaging these bundles. Tractography is basically a non-invasive method for visualization of the white matter fiber bundles or tracts done using information in diffusion magnetic resonance images (dMRI) (YAMADA *et al.*, 2009). Recently, machine learning approaches and deep-learning-based methods have been used for tractography (NEHER *et al.*, 2017; POULIN *et al.*, 2017). However, up to date, none have investigated CNN for such task.

1.2 Objectives

In this dissertation, we advance the study of deep-learning-based approaches for brain MR analysis. Two problems were tackled using deep-learning-based approaches for brain MR images. First, we investigated this technique in a segmentation task for skull-stripping (SS) and then in a regression analysis for tractography.

1.3 Contributions

The main contributions of this work are in the deep-learning-based segmentation and regression domains in brain MR. Two specific tasks were considered: skull-stripping (SS) and tractography. The main contributions of this work are summarized below.

1.3.1 Skull-stripping

To begin with, we developed an end-to-end CNN-based solution using what we refer to as "silver-standard" annotation for the brain extraction or skull-stripping. Segmenting brain tissue from non-brain tissue is a process known as brain extraction or SS. The silver standard annotations are masks generated by an agreement among eight publicly available automatic SS methods using the consensus algorithm Simultaneous Truth and Performance Level Estimation (WARFIELD *et al.*, 2004). An outline of our findings for this task is presented as follows:

- Validation of silver standard masks for CNNs training.

- Development of a deep-learning-based method for SS fully trained with consensus-derived silver standard masks, eliminating the use of expert manual annotation.
- Cross-dataset validation for deep-learning-based SS approach.
- State-of-the-art performance.
- Open-source and publicly available approach.

1.3.2 Tractography

As a second application, we investigated a CNN-based-tractography solution for epilepsy surgery, which was done during a six-month internship at University College London under the FAPESP BEPE scholarship. The main goal of this analysis was to structure a baseline for a deep-learning-based-regression to predict white matter fiber orientations. We outline our main findings as follows:

- A baseline using deep-learning-based regression for tractography.
- Investigation of convolutional neural networks for tractography.
- Robust regression using a three stages approach with an auto-encoder, training from scratch, and fine-tuning.

1.4 Organization of the Dissertation

This dissertation is organized into three main parts, which are: Convolutional Neural Networks for skull-stripping in Brain MR Imaging using Consensus-based Silver standard Masks (Chapter 2), Deep-learning-based Tractography for Surgical Planning in Epilepsy Treatment (Chapter 3), and Conclusions (Chapter 4). Remark: both Chapter 2 and 3 are written in an article-like style.

In Chapter 2, we detail the current state-of-the-art and challenges, our contributions, the datasets we adopted, the steps for the generation of our silver-standard masks, our proposed pipeline, our results, discussion, and finally the conclusions.

In Chapter 3, we follow the same organization as the previous one except for not detailing of silver standard masks since they are not used for this case.

In Chapter 4, we present our final thoughts regarding the works presented in Chapters 2 and 3 and a list of publications related to this M.Sc. research.

2 Convolutional Neural Networks for Skull-stripping in Brain MR Imaging using Consensus-based Silver standard Masks

2.1 Motivation

Segmenting brain tissue from non-brain tissue (a process known as brain extraction or skull-stripping, SS) is a critical step in many MR brain image processing algorithms. After brain extraction, the analysis of brain regions are more easily and more accurately performed (KALAVATHI; PRASATH, 2016), thus, accurate brain segmentation is an essential, early processing step. In fact, it is typically the initial step in a wide range of brain MR imaging analyses, such as when segmenting tissue types (BOER *et al.*, 2010), analyzing multiple sclerosis lesions (ZIVADINOV *et al.*, 2004), classifying Alzheimer's disease (RUSINEK *et al.*, 1991), assessing schizophrenia (TANSKANEN *et al.*, 2005), monitoring the development or aging of the brain (BLANTON *et al.*, 2004), and determining changes in volumes and shape of brain regions across many disorders (PETRELLA *et al.*, 2003; HUTCHINSON; RAFF, 2000). Normally, brain MR images present unwanted non-brain tissues that make SS challenging. Further, the brain gyri and sulci (*i.e.*, the ridges and depression on the brain outer surface, respectively) can challenge even current state-of-the-art SS methods (IGLESIAS *et al.*, 2011). New approaches are continually being proposed to overcome these and other limitations, suggesting that the study of SS techniques remains an active research field using either conventional methods (IGLESIAS *et al.*, 2011; SMITH, 2002; SÉGONNE *et al.*, 2004; ESKILDSEN *et al.*, 2012; BEARE *et al.*, 2013; AVANTS *et al.*, 2011; SHATTUCK *et al.*, 2001) or, more recently, deep learning (DL)-based approaches (KLEESIEK *et al.*, 2016; SALEHI *et al.*, 2017).

After the groundbreaking result of (KRIZHEVSKY *et al.*, 2012), DL, especially CNN-based approaches, has become a commonly employed algorithmic approach to solve medical imaging problems (LITJENS *et al.*, 2017). DL-based methods are trained with labeled raw data to "automatically discover" the underlying mathematical representations needed for detection, classification and/or segmentation (LECUN *et al.*, 2015). Commonly, training a CNN from scratch requires a large amount of correctly labeled data. Appropriate medical image datasets, however, are generally too small to succeed at this task. Challenges often arise because labeled data require significant manual effort from an expert in order to complete this time-consuming and, thus, expensive task (GREENSPAN *et al.*, 2016).

To reduce cost, single rater, manual annotation is often used. However, manual annotation is known to vary, even among highly-trained experts (WARFIELD *et al.*, 2004; AKKUS *et al.*, 2017), and be impacted by both inter- and intra-rater variability (ASMAN; LANDMAN, 2011). Additionally, the characteristics of MR data are complex and can be impacted by several other factors including contrast differences among scanners and changes in image spatial resolution, especially at border voxels lying between tissues. These and other issues lead to the presence of ambiguous voxels and label confusion during manual annotation by experts (ASMAN; LANDMAN, 2011). Finally, manual annotation guidelines are generally subjective. For example, there are more than a dozen protocols for hippocampal segmentation and different protocols have been shown to provide up to 2.5-fold variation in volume estimates (BOCCARDI *et al.*, 2011). An effective scenario to mitigate the inherent variability in manual annotation is development of a consensus agreement approach that uses multiple expert annotations in order to generate a robust "gold standard". However for tasks like SS, forming a consensus among multiple experts is impractical due to the linear increase in cost associated with performing each additional manual segmentation. Consensus-based approaches can also generate annotated data by finding agreement between different annotations or between the outputs of different automated methods (WARFIELD *et al.*, 2004; ASMAN; LANDMAN, 2011; REX *et al.*, 2004). These consensus results are potentially robust and in the past have been applied to improve automatic multi-atlas segmentation methods (ALJABAR *et al.*, 2009; WU *et al.*, 2014).

2.1.1 Previous Works in Skull Stripping

Traditional (*i.e.*, non DL-based) SS methods can be categorized into one of six main classes: 1) manual annotation, 2) intensity-based methods, 3) morphology-based methods, 4) deformable surface-based methods, 5) atlas-based methods, and 6) hybrid methods (KALAVATHI; PRASATH, 2016). The gold standard method, manual annotation, is usually done by an expert, often a radiologist or similarly highly trained user. Manual methods, unfortunately, are time consuming; experts often need to spend hours segmenting one brain image volume. Manual annotations are considered, thus, impractical for medical analysis in large-scale studies. The second class, intensity-based methods (SHATTUCK *et al.*, 2001), are fast, but lack robustness. They are very sensitive to local changes in image contrast, noise, and artifacts. Morphology-based methods (BEARE *et al.*, 2013), the third class, are also fast, but depend on parameters that are experimentally computed and related to size and shape of mathematical morphological operations (KALAVATHI; PRASATH, 2016). Deformable surface model-based methods (SMITH, 2002) are a class that use a balloon-like template that deforms to fit the brain based on gradient information. Although they can fit both the interior and exterior areas of the brain, these

methods are very dependent on initialization of the balloon-like template. Atlas-based methods (ESKILDSEN *et al.*, 2012; AVANTS *et al.*, 2011) rely on image registration to an atlas template, making them time-consuming approaches and very dependent of the atlas geometry. Lastly, hybrid methods attempt to combine the best features of the previously described methods. They generally require longer processing times, but usually achieve optimal segmentation results (IGLESIAS *et al.*, 2011; SÉGONNE *et al.*, 2004; LUTKENHOFF *et al.*, 2014).

DL-based segmentation are newer methods that are performed using two predominant approaches: 1) voxel-wise networks using CNN architectures with fully connected layers that classify the central pixel in an image patch, and 2) fully convolutional networks (FCNs) (LONG *et al.*, 2015) that segment the entire image in one step. Both methods have been implemented using both 2D and 3D architectures; but because 3D convolutions are computationally expensive, 2D convolutions are more commonly used. Although the first class of approaches have been frequently exploited due to their derivation from classification tasks, FCNs perform better in retrieving spatial information from local and global features. They are also faster than voxel-wise networks (LONG *et al.*, 2015; RONNEBERGER *et al.*, 2015). Moreover, FCNs can work with any sized input because their weights do not depend on the input size, a limitation of voxel-wise networks.

Recently, two DL-based SS methods have been proposed; both methods were validated against small publicly available datasets against manual annotation. (KLEESIEK *et al.*, 2016) proposed a voxel-wise 3D CNN for SS that we will refer to as 3D CNN. The 3D CNN is not deep due to the cost of the 3D convolutions, limiting its learning capacity. (SALEHI *et al.*, 2017) applied the auto-net method to brain extraction. They examined two approaches: 1) parallel voxel-wise networks and 2) parallel 2D FCN U-Net (RONNEBERGER *et al.*, 2015), each followed by an auto-context CNN classifier. Basically, this classifier takes the concatenation of the probability maps from a pre-trained network and feeds them as input data to another CNN following the auto-context algorithm presented in (TU; BAI, 2010). Context information retrieval with CNNs have been explored in other medical segmentation tasks. (CHEN *et al.*, 2017) also presented an auto-context version of its VoxResNet approach, which is an architecture based on the ResNet (HE *et al.*, 2016) for brain structures segmentation. (KAMNITSAS *et al.*, 2017) used multi-scale 3D CNN with fully connected conditional random fields (CRFs) for brain lesion segmentation, but such CRFs are time consuming and have very limited neighborhood relations compared to the auto-context approach (TU; BAI, 2010).

Consensus methods could be used to generate annotated data from the output masks of different automatic methods. (REX *et al.*, 2004) obtained a higher agreement rate than that of individual segmentations done by two different experts. Recently, consensus masks have been used to generate what we refer to as silver standard masks.

(SOUZA *et al.*, 2017) evaluated the agreement between consensus predictions and manual labeled data in the *Calgary-Campinas-359* (*CC-359*) public dataset in which silver standard masks were generated by the consensus algorithm simultaneous truth and performance level estimation (STAPLE) (WARFIELD *et al.*, 2004). This work also suggested the usage of consensus masks for training CNNs. (LUCENA *et al.*, 2018) have further investigated and validated the usage of silver standard masks in the CNN training stage for SS. In this work, the authors trained and compared the same DL architecture with silver standard annotation labels and manual annotation labels. Their results suggested that the performance of training a network with silver standard labeled data are comparable to models trained with gold standard data but generalize better due to consensus method, likewise STAPLE, because it reduces the inter-/intra-rater variability. Also, they can be generated without the need for (and cost of) manual annotation, potentially augmenting our training input datasets and improving generalization over training a CNN with only a single manual annotation.

2.1.2 Our Approach

We present in this chapter a CNN approach for brain extraction in MR images. Unlike others, our method is completely trained using silver standard masks that are generated by forming the consensus between eight, public, non DL-based automatic SS methods. Our method has two main implementations: 1) a tri-planar method using parallel 2D CNNs that we will refer to as CONSNet and 2) an auto-context variation of CONSNet that adapts an auto-context CNN in cascade with the tri-planar method. The term CONSNet is used to refer to the complete approach (*i.e.*, training with silver standard masks and use of a CNN architecture). Our analysis were conducted on three public datasets: *Calgary-Campinas-359* (*CC-359*) (SOUZA *et al.*, 2017), LONI Probabilistic Brain Atlas (LPBA40) (SHATTUCK *et al.*, 2008), and the Open Access Series of Imaging Studies (OASIS) (MARCUS *et al.*, 2007). We validated our method against manual annotations available in twelve image sets in the *CC-359* dataset and all image sets in the LPBA40 and OASIS datasets. Five performance metrics were used: Dice coefficient, sensitivity, specificity, Hausdorff distance, and symmetric surface-to-surface mean distance. Furthermore, we compared the processing time of our method against the publicly available state-of-the-art automatic SS methods and the consensus-based masks generated by STAPLE.

The main contributions of our approach can be summarized as follows:

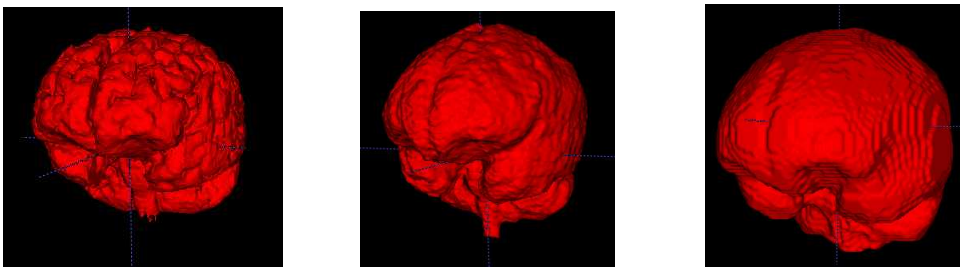
- It is the first DL-based method to be fully trained with consensus-derived silver standard masks. This step eliminates the cost associated with manual annotation. It also enlarges the training input data so that it is suitable for large-scale analysis from non-annotated data.

- It is generalizable. It was trained using the *CC-359* dataset and was then validated using the LPBA40 and OASIS datasets.
- It outperforms most SS methods including some DL-based approaches.
- It is completely open-source and publicly available[1].

2.2 Materials and Methods

2.2.1 Datasets

Three publicly available datasets were used for this study. All datasets contain adult human MR brain images acquired using a T1-weighted volumetric imaging method. Some variability in the image acquisition parameters is presented within and between the three datasets. The *CC-359* was used for training and both the LPBA40 and OASIS dataset were used for validation. The three public datsets have a total of 476 subjects (218 males, 258 females, 51.2 ± 10.4 years). All three datasets included manual (gold standard) segmentations of varying quality (Figure 1).



(a) Manual *CC-359* (b) Manual LPBA40 (c) Manual OASIS

Figure 1 – Representative 3D reconstruction of the manual (gold standard) annotation for one subject of the *CC-359*, OASIS and LPBA40 datasets.

2.2.1.1 *CC-359* Dataset

The *CC-359*[2] is a public dataset composed of image volumes in NIfTI format from 359 subjects (176 males, 183 females, 53.5 ± 7.8 years) acquired in the coronal image plane. Data were collected on scanners from three different vendors (General Electric (GE) Healthcare, Philips Medical Systems, and Siemens Healthineers) and at two magnetic field strengths (1.5 T and 3 T). Image volumes have a spatial resolution of $1.0 \times 1.0 \times 1.0$ mm^3 (SOUZA *et al.*, 2017). The *CC-359* dataset includes the original volumes, the consensus masks were generated for all subjects using the STAPLE algorithm (described

[1] https://github.com/MICLab-Unicamp/CONSNet
[2] http://miclab.fee.unicamp.br/tools

in Section 2.2.3). In addition, manual annotations were performed on twelve randomly selected subjects (two for each vendor-field strength combination).

2.2.1.2 LPBA40 Dataset

The LPBA40[3] dataset is composed of 40 T1-weighted image volumes from healthy subjects (20 males, 20 females, 29.2 ± 6.3 years) and their corresponding manually labeled brain masks (SHATTUCK *et al.*, 2008). The scans were acquired on a GE Healthcare 1.5-T system with a spatial resolution of $0.86 \times 1.5 \times 0.86$ mm^3.

2.2.1.3 OASIS Dataset

We use the first two discs of the OASIS[4] dataset that consist of T1-weighted volumes from 77 subjects, with spatial resolution of $1.0 \times 1.0 \times 1.0$ mm^3 (MARCUS *et al.*, 2007). This dataset contains 55 females and 22 males with an average age of 51.6 ± 24.7 years. Twenty subjects (26%) had early Alzheimer's disease. The masks in this dataset were segmented with a custom method based on registration to an atlas, and then revised by human experts (IGLESIAS *et al.*, 2011); unlike *CC-359* and LBPA40, the OASIS images were not fully manually segmented. As a result, the quality of the masks provided with this dataset is relatively poor (*cf.*, Figure 1c) - cortical surface features and other small structures are filtered. Nonetheless, we choose to use this dataset so that we can compare our results against published results of (KLEESIEK *et al.*, 2016) and (SALEHI *et al.*, 2017).

2.2.2 Automatic Skull-stripping Methods

This work employed a series of eight state-of-the-art, non DL-based, automatic skull-stripping methods, as well as two DL-based methods. These methods are all public and were used to develop consensus-derived labeled data (non DL-based methods only, Section 2.2.3) and to analyze the performance of our proposed SS methods (Section 2.2.6). In alphabetical order, the eight non DL-based methods were: 1) Advanced Normalization Tools (ANTs) (AVANTS *et al.*, 2011), 2) Brain Extraction based on non-local Segmentation Technique (BEAST) (ESKILDSEN *et al.*, 2012), 3) Brain Extraction Tool (BET) (SMITH, 2002), 4) Brain Surface Extractor (BSE) (SHATTUCK *et al.*, 2001), 5) Hybrid Watershed Approach (HWA) (SÉGONNE *et al.*, 2004), 6) Marker Based Watershed Scalper (MBWSS) (BEARE *et al.*, 2013), 7) Optimized Brain Extraction (OPTIBET) (LUTKENHOFF *et al.*, 2014), and 8) Robust Brain Extraction (ROBEX) (IGLESIAS *et al.*, 2011). An overview of non DL-based methods was provided in motivation

[3] http://www.loni.usc.edu/atlases
[4] http://www.oasis-brains.org/app/template/Index.vm

section and further details can be found in the cited references. In our analysis we used the default processing parameters detailed in the above citations.

The two DL-based methods are the 3D CNN (KLEESIEK *et al.*, 2016) and auto-net (SALEHI *et al.*, 2017). The 3D CNN approach is a voxel-wise network containing seven convolutional hidden layers, and one convolutional soft-max output-layer. The receptive field (*i.e.*, input for each predicted pixel) of this model is 53^3 voxels (KLEESIEK *et al.*, 2016). For this method, we used the CNN model provided by the authors, which was trained using three public datasets (LPBA40, Internet Brain Segmentation Repository (IBSR), and OASIS). We tried to train the 3D CNN using our data, but the results were worse than the published model. Therefore, the published model was used in this work. The results of a two-fold cross-validation experiment with the LPBA40 and OASIS datasets were used in our comparative analysis.

A second DL-based approach used auto-net: a 2D FCN U-Net followed by an auto-context CNN classifier. (SALEHI *et al.*, 2017) describe their results with and without the auto-context CNN in a two-fold cross-validation experiment using the LPBA40 and OASIS datasets. Only these published results were used as the authors did not provide source code.

2.2.3 STAPLE-derived Silver Standard Consensus

Consensus methods can be used to provide more reliable and accurate segmentation labeling in SS and other image processing tasks. These methods combine different segmentations and obtain more robust results (WARFIELD *et al.*, 2004; ASMAN; LANDMAN, 2011; REX *et al.*, 2004; REHM *et al.*, 2004). STAPLE is one such consensus-forming algorithm that uses an expectation-maximization algorithm to estimate the hidden (or true) segmentation as a probabilistic mask. The algorithm considers a collection of segmentations and computes a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation method (WARFIELD *et al.*, 2004). This algorithm was used in this work to generate what we refer to as silver standard segmentation masks. These brain masks are formed from STAPLE output (a probability mask) by applying a threshold of 0.5. In our study, STAPLE used as an input binary masks resulting from the eight non DL-based automatic skull-stripping techniques previously described (Section 2.2.2).

We applied the STAPLE algorithm to the LBPA40 and OASIS dataset (It has already been applied to the *CC-359* dataset using the same protocols that we adopted.) STAPLE was chosen in this work because the algorithm has been validated extensively through experiments (WARFIELD *et al.*, 2004; FENNEMA-NOTESTINE *et al.*, 2006; CAN *et al.*, 2009). Moreover, an open-source implementation of the algorithm was available (Insight Segmentation and Registration Toolkit (ITK) repository (JOHNSON *et al.*,

2015)). For the *CC-359* dataset, silver standard masks for the *CC-347* subset were generated and used for CNN training. The consensus brain masks generated for the twelve subjects with gold standard manual annotation (*i.e.*, the *CC-12* subset) were compared against the manual annotations. For clarity, the silver standard masks derived from the *CC-12* are referred to as STAPLE-12. For LPBA40 and OASIS datasets, the silver standard masks were only compared against manually annotated data. These silver standards are referred to as STAPLE-LPBA40 and STAPLE-OASIS.

2.2.4 Convolutional Neural Network Architecture and Implementation

Our CNN is based on the 2D FCN U-Net architecture. The original U-Net architecture is a "U"-shaped network (contracting path, left side; expansive path, right side, Figure 2) composed of 23 convolutional layers (RONNEBERGER *et al.*, 2015). Our implementation is a modification of the CNN architecture from RECOD Titans (MENEGOLA *et al.*, 2017) which is composed of 20 convolutional layers.

In our implementation, we removed the fully connected layers and used a fixed kernel size of 3×3. We adopted the RMSprop (TIELEMAN; HINTON, 2012) with an initial learning rate of 0.001 and an exponential decay of 0.995 after each epoch at the training stage. Additionally, the negative of the Dice coefficient (described in Section 2.2.6) was used as the loss function. The implementation was built using Keras with Tensorflow (ABADI *et al.*, 2016) as a backend. The full code of our implementation is openly available at https://github.com/MICLab-Unicamp/CONSNet.

2.2.5 Proposed Brain Extraction Pipelines

We propose two DL-based brain extraction methods: 1) the regular CONSNet that consists of a tri-planar method, analogous to what was implemented in (PRASOON *et al.*, 2013), with three parallel 2D CNN pipelines and 2) an auto-context version of CONSNet. Both methods are summarized pictorially in Figure 3. The pipeline presented in Figure 3a shows the steps required to perform the CONSNet prediction where its final output is calculated after applying a threshold to the average probability of the three CNN output probability maps. The pipeline presented in Figure 3b overviews the auto-context CONSNet version. This second version takes the output probability maps from the three CNNs as the input to a fourth CNN.

The key idea in our implementation is to perform 2D segmentation on a image-by-image basis over each volume and repeat for each orthogonal orientation (*i.e.*, original axial plus both the coronal and sagittal reformatted images). 3D segmentation is then done by reconstruction through the concatenation of the 2D predictions. Our approach is analogous to what is done manually by an expert when reviewing volumetric images.

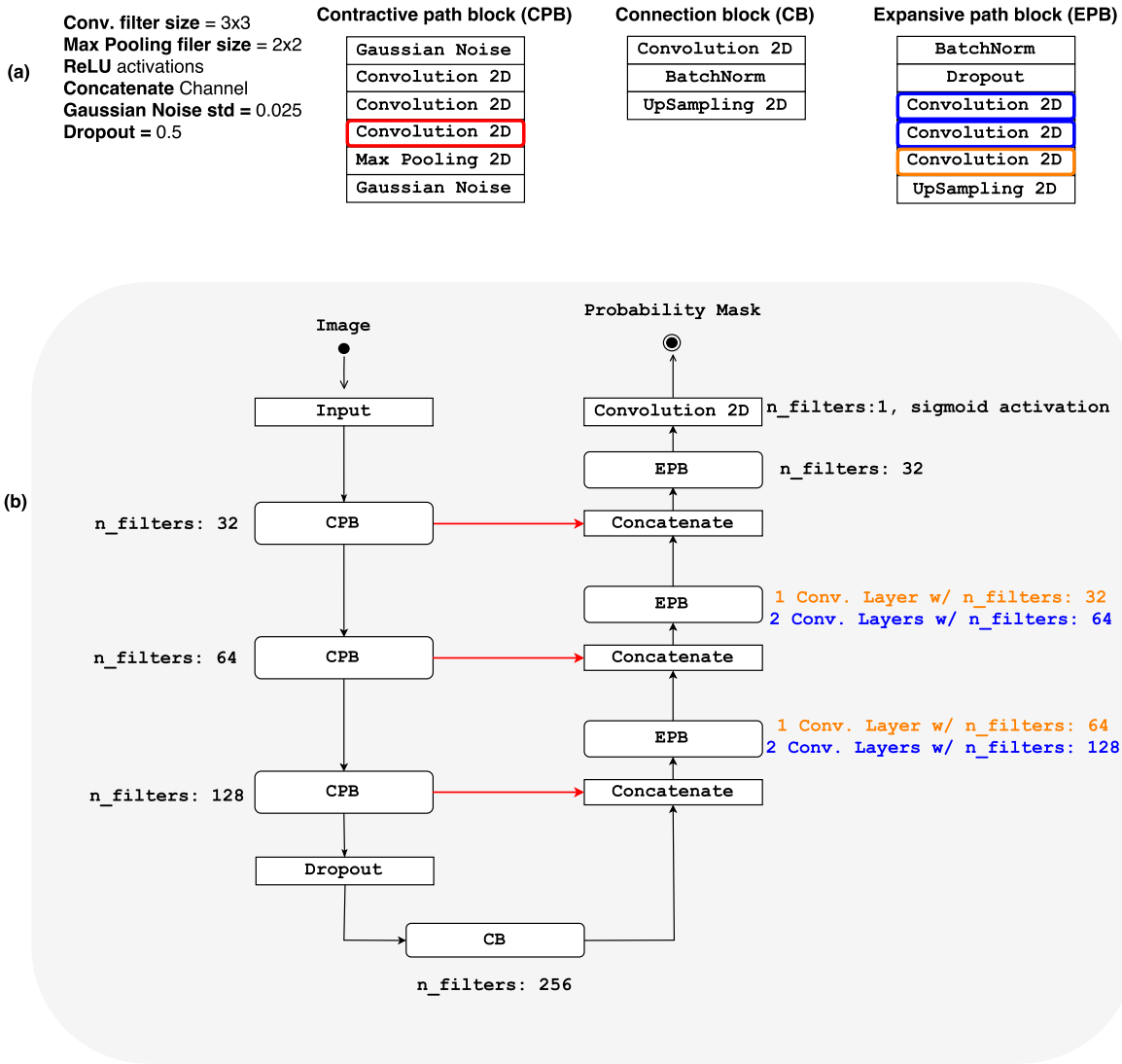


Figure 2 – Modified RECOD U-Net architecture. (a) presents configurations of the contractive path blocks (CPB), connection blocks (CB), and expansive path blocks (EPB) used in the architecture. (b) illustrates the entire architecture consisting of CPB, CB, and EPB modules. The contracting path is on the left and the expansive path is on the right. The concatenations (red arrows) are always done between the output of the third convolutional layer (red in CPB block of (a)) in the contracting path and the output of the previous block in the expansive path. The number of filters in two first EPB blocks vary for each convolution. The text in blue and orange of (b) correspond to the convolution layers blocks of the same colors in the EPB block in (a).

To manually segment an image volume, a human operator would start the task in one view (often sagittal), and toggle to the other orthogonal views (axial and coronal) to determine that the voxel in question is or is not brain tissue. The operator is undertaking a form of “label voting" (CHAKRAVARTY *et al.*, 2013) by assessing each voxel from different perspectives. By adopting the tri-planar approach, we expect that the CONSNet

prediction improves compared to when only processing one orientation.

The auto-context CONSNet implementation takes the probability maps generated by 2D parallel CNNs as input to a fourth CNN. The algorithm refines the output results in an iterative way and integrates low-level and contextual information by fusing a large number of low-level appearance features with context and implicit shape information (CHEN *et al.*, 2017). Originally (TU; BAI, 2010) used random forest classifiers to perform this step; here we take the advantage of the CNN architecture to extract the context information.

In both implementations, CONSNet and auto-context CONSNet, there are three major steps: 1) a pre-processing step, 2) a CNN segmentation step, and 3) a threshold and post-processing step. The details of each step are described in the following subsections.

2.2.5.1 Pre-processing Step

The *CC-347* data subset, which is composed of 347 volumes and their respective silver standard masks, was used to train our CNNs. Because the range of grayscale intensities varies across the image volumes, we first normalized each volume to the same image intensity range (0 to 1000). This range was chosen to ensure sufficient dynamic range and to minimize data storage requirements.

Second, because the volumes from different vendors have differing matrix dimensions, we initially varied the number of patches and their size to improve spatial content retrieval. We settled on using three patches of size 128×128 per slice in each image volume. Patches were randomly extracted from each slice that contained brain voxels, and they were subsequently fed into the CNN during training.

We did not use patches surrounding a common voxel; rather, patches were randomly extracted across each slice. This approach was applied in each axial, coronal and sagittal image orientation for the CONSNet training and after concatenation of the ouput probability maps for the auto-context CONSNet training.

2.2.5.2 CNN Segmentation Step

In the training step for CONSNet (Figure 3a), the three 2D parallel CNNs, one for each axial, coronal, and sagittal image plane, were trained. The input to the final prediction step consisted of the predictions of all three orthogonal models. For the auto-context CONSNet (Figure 3b), the training step consisted of training a fourth CNN using three channels (*i.e.*, one channel per each tri-planar model prediction). This step was needed because we have to concatenate the output probability maps from the other CONSNet steps to produce an input to the fourth network. For this implementation, our CNN uses the sagittal orientation.

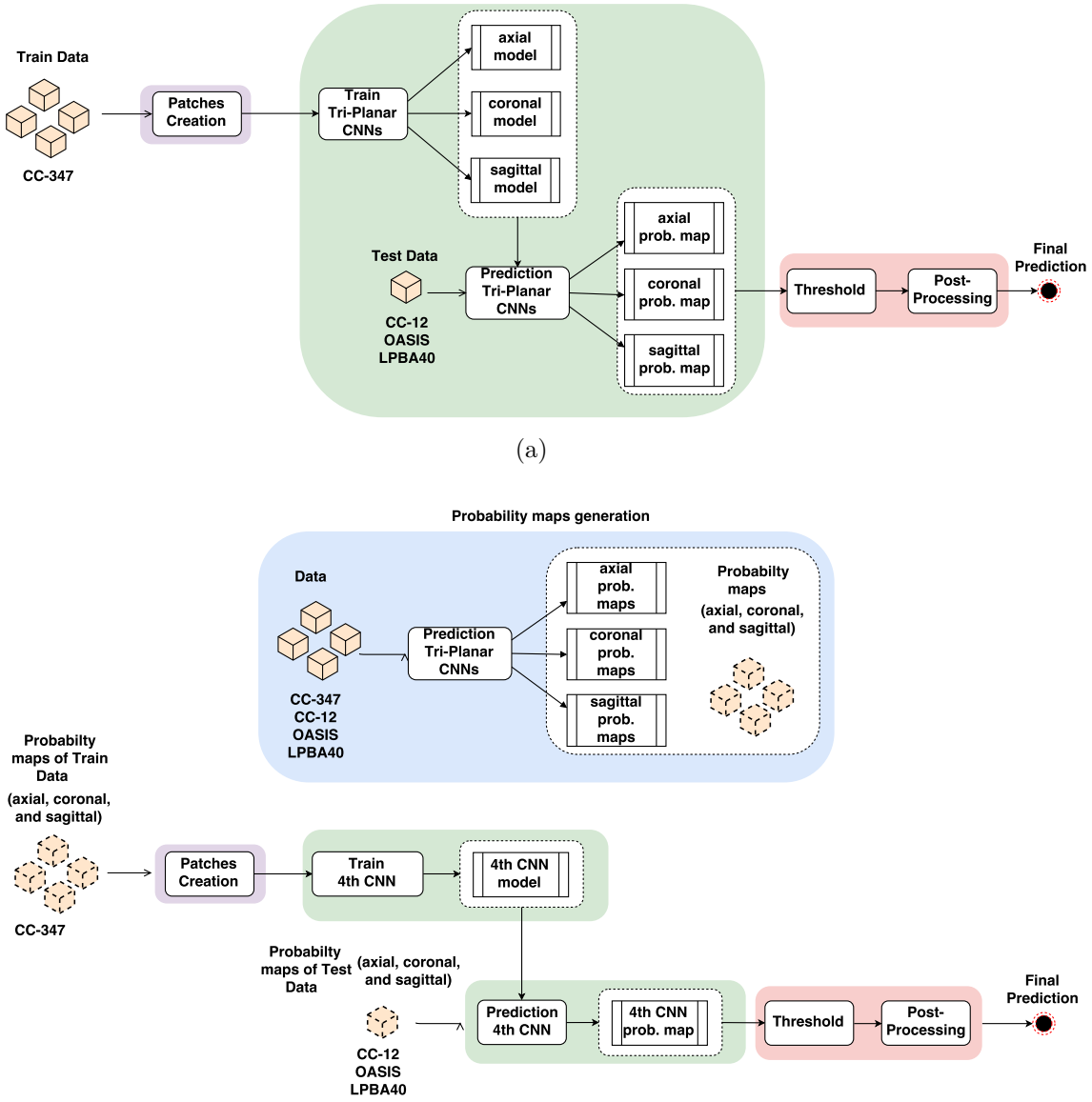


Figure 3 – Proposed deep learning brain segmentation pipelines. Both pipelines consist of three stages: pre-processing (purple), CNN segmentation (green), and threshold/post-processing (red). The CONSNet pipeline is shown in (a), and the auto-context CONSNet pipeline is shown in (b). The blue box in Figure (b) represents the probability generation done by CONSNet to be used as input in auto-context CONSNet.

We chose this plane because it is the plane most often considered by experts when performing manual annotation. As previously mentioned, experts generally toggle to the orthogonal axial and coronal views only to ensure that a voxel is or is not brain tissue.

2.2.5.3 Threshold and Post-processing Step

In the first CONSNet pipeline (Figure 3a), each of the CNN models considers different orientation-specific spatial information. Therefore, to retrieve integrated spatial information from the predictions provided by the three CNNs, we calculated the average probability of the results, and then thresholded the result. The threshold consisted of setting to 1.0 voxels where the average probability from the three CNNs was greater than 0.5. Otherwise, the voxels were set to 0.0. The second CONSNet pipeline (Figure 3b) produced only one prediction, thus, the threshold was applied directly to the single CNN output. A threshold of 0.5 was applied to produce a final prediction. Both the CONSNet and auto-context CONSNet final predictions were obtained after a post-processing step in which only the largest connected component was preserved. The other smaller components were removed using an area-open (SALEMBIER *et al.*, 1998) filter.

2.2.6 Evaluation Metrics and Statistical Analysis

The evaluation metrics used were: sensitivity, specificity, Dice coefficient, Hausdorff distance, and mean symmetric surface-to-surface distance. The sensitivity, specificity and Dice coefficient metrics are overlap metrics (larger numbers are best, maximum value is 100%). The Hausdorff distance and mean symmetric surface-to-surface distance represent surface distance metrics (smaller numbers are best, minimum value is 0.0). If we let G be the gold standard image and S the segmentation we wish to compare, then the metrics can be defined by:

- Dice coefficient:
$$Dice(G,S) = \frac{2TP}{2TP + FP + FN}$$
- Sensitivity:
$$Sensitivity(G,S) = \frac{TP}{TP + FN}$$
- Specificity:
$$Specificity(G,S) = \frac{TN}{TN + FP}$$
- Hausdorff distance:
$$d_H(S,G) = \max\{\sup_{s\in S}\inf_{g\in G} d(s,g), \sup_{g\in G}\inf_{s\in S} d(s,g)\}$$
- Symmetric surface-to-surface mean distance:
$$d_S(S,G) = \frac{\sum_{s\in S}\min_{g\in G} d(s,g) + \sum_{g\in G}\min_{s\in S} d(g,s)}{|S| + |G|}$$

where TP, FP, TN, and FN are the number of true positive, false positive, true negative, and false negative findings, respectively, $d(\cdot, \cdot)$ is the Euclidean distance, sup is supremum, and inf is the infimum.

Each adopted metric evaluates different performance characteristics between the prediction (S) and ground truth (G) segmentations. Sensitivity measures how much brain tissue is included in the segmentation, whereas specificity measures how much non-brain tissue is correctly segmented as non-brain. The Dice coefficient metric is a compromise between sensitivity and specificity metrics; it evaluates the trade-off between the correct and incorrect voxel classifications. The Hausdorff distance is indicative of segmentation outliers, and the symmetric surface-to-surface mean distance has a similar interpretation to the Dice coefficient but uses the distance between the segmented and gold standard surfaces.

Heat maps were created to better visualize the correctness of the segmented voxels versus the manual mask. Non-linear registration (AVANTS *et al.*, 2011) was used to place all subjects on the same space. The average FP and FN error was then projected for all the skull stripping methods. The manually segmented subjects served as the reference and sagittal, coronal, and axial projections were formed.The projected values were normalized between 0.0 and 1.0.

The statistical analysis to assess differences in the evaluation metrics was done using Wilcoxon signed-rank tests with Bonferroni correction. This test is a non-parametric statistical hypothesis test that does not assume a normal distribution (HAYNES, 2013). A p-value less than 0.05 was used to confirm statistically significant differences. For purposes of statistical comparison, the auto-context CONSNet approach was selected as the reference method.

2.2.7 Experimental Methodology

CONSNet was compared, in two experiments against the eight, traditional (non DL-based) methods, the STAPLE-derived consensus of these methods, and two DL-based methods (3D CNN (KLEESIEK *et al.*, 2016) and auto-net and U-Net (SALEHI *et al.*, 2017)). In Experiment 1, we trained CONSNet and the auto-context CONSNet using the STAPLE-derived consensus formed from the *CC-347* subset to demonstrate the concept of training using silver standard data and the model generalizability. In Experiment 2, we demonstrated the robustness of our CNN architecture by performing a two-fold cross-validation using the LPBA40 and OASIS datasets. The labeled masks for this experiment are the ones originally provided by LPBA40 and OASIS datasets. Each experiment is described in the following sections. All experiments used a general purpose workstation equipped with a CPU (Xeon® E3-1220 v3, 4×3.10 GHz; Intel) and 32 GByte of memory. The workstation GPU (GeForce Titan X; NVIDIA) had 12 GByte of on-board memory.

2.2.7.1 Experiment 1

In Experiment 1, we demonstrated training of the CONSNet and the auto-context CONSNet methods using the STAPLE-generated silver standard masks derived from the *CC-347* data subset. We compared our findings against the eight non DL-based SS methods, plus the 3D CNN (KLEESIEK *et al.*, 2016) method and the STAPLE consensus results for the *CC-12*, OASIS and LPBA40 datasets.

2.2.7.2 Experiment 2

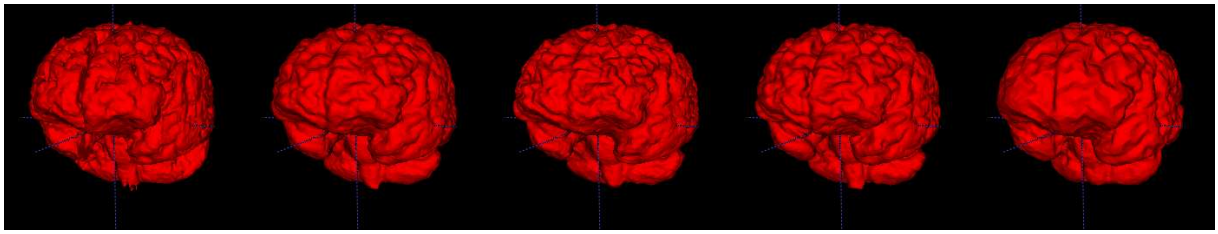
In the second experiment, we performed a two-fold cross-validation using separately the LPBA40 and OASIS datasets. These tests compare our findings against published results for 3D CNN (KLEESIEK *et al.*, 2016), and for auto-net and U-Net (SALEHI *et al.*, 2017) networks. We also evaluated model generalization by validating our approaches using the *CC-12* subset and either the OASIS or LPBA40 dataset (*i.e.*, excluding the dataset used for network training). In Experiment 2, the patch size was allowed to vary; it was experimentally set, allowing the patch size to change according to the image size of each dataset. We used ten patches of size 96×96 for the LPBA40 dataset and ten patches of size 144×144 for the OASIS dataset.

2.3 Results

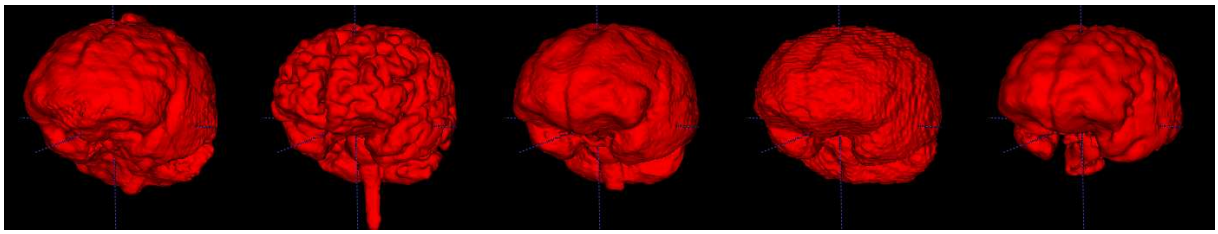
2.3.1 Experiment 1

Figures 4 to 6 present representative 3D reconstructions of the different segmentation methods for one sample subject in the *CC-12*, OASIS, and LPBA40 datasets, respectively. These panels compare the manual annotation to the outputs of both the published non DL and DL-based methods, and to our proposed CONSNet methods. Overall the quality of the brain extraction is good, though some variation by method and by dataset was found. For instance, there are automatic methods that poorly delineate (or smooth) the gyri and sulci, such as ROBEX (*cf.*, Figure 4a *vs* Figure 4i). The OASIS manual annotation (Figure 6a) is also relatively smooth because it was automatically segmented using a computer and then manually revised (see description provided in Section 2.2.1.3). The BSE method was found to perform much worse than the other methods with the *CC-12* data subset. Therefore, we chose to not include the BSE output masks in our subsequent Experiment 1 analyses. To generate the STAPLE consensus results, however, we included the BSE image volumes.

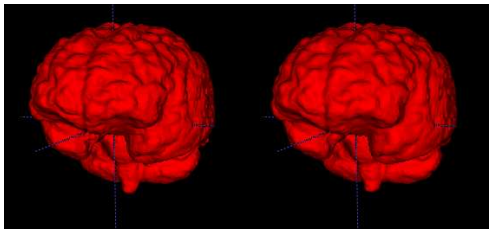
Tables 1 to 3 summarize the overall analyses for each evaluation metric. These data are summarized as box plots in Figures 7 to 11. Performance of the CONSNet method was excellent across the three datasets. While all tested methods generally performed



(a) Manual (b) STAPLE-12 $prob > 0.5$ (c) ANTs (d) BEAST (e) BET



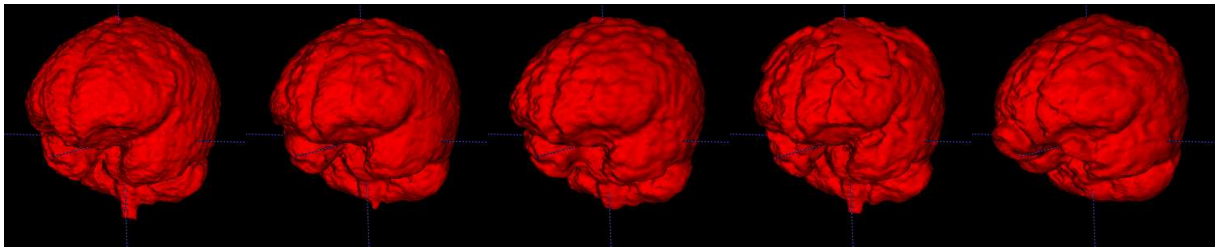
(f) HWA (g) MBWSS (h) OPTIBET (i) ROBEX (j) 3D CNN



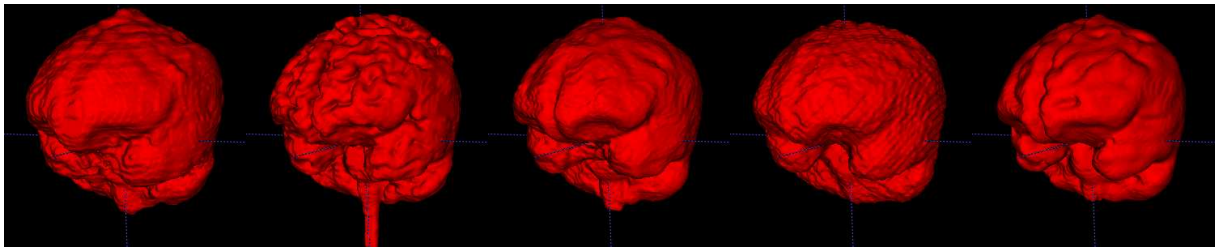
(k) CONSNet (l) auto-context CONSNet

Figure 4 – Representative 3D reconstruction of the different segmentation methods for one subject of the *CC-12* subset.

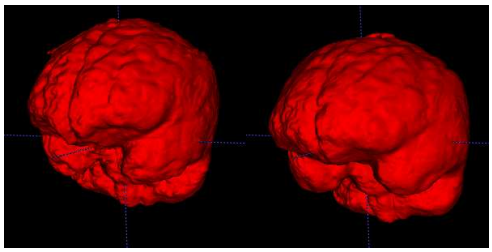
well (*i.e.*, $> 90\%$ on Dice coefficient, sensitivity and specificity; < 10 mm on Hausdorff distance; and < 0.1 mm on symmetric surface-to-surface mean distance), CONSNet was typically in the top five of the eleven tested methods (except for specificity). In the OASIS dataset (Table 3), our CONSNet method performed slightly worse than ROBEX, however as previously noted, the ROBEX segmentations were very smooth and did not track surface brain tissue details. Finally, CONSNet outperformed the 3D CNN method (Dice coefficient, sensitivity, Hausdorff and symmetric surface-to-surface mean distances) across all three datasets. The results from the LPBA40 and OASIS datasets suggest generalizability of the *CC-347*-trained network.



(a) Manual (b) STAPLE-LPBA40 $prob > 0.5$ (c) ANTs (d) BEAST (e) BET



(f) HWA (g) MBWSS (h) OPTIBET (i) ROBEX (j) 3D CNN

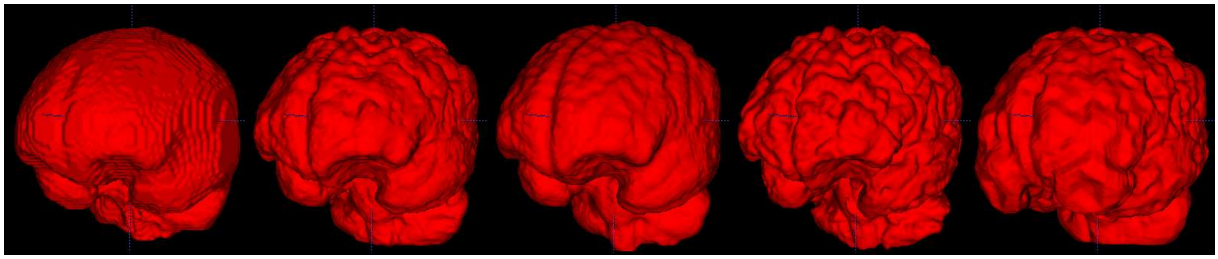


(k) CONSNet (l) auto-context CONSNet

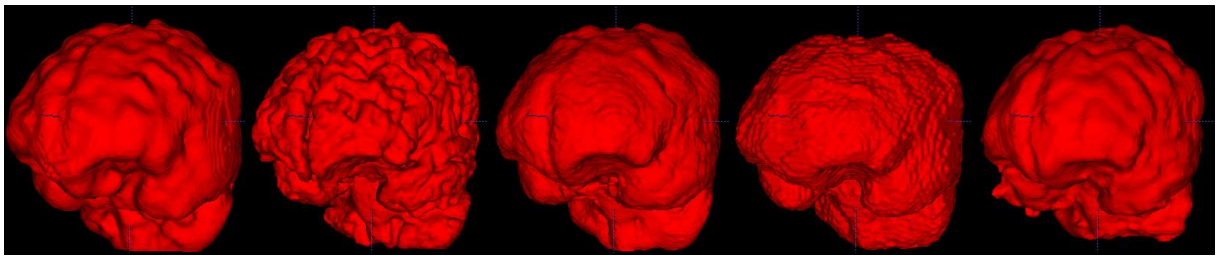
Figure 5 – Representative 3D reconstruction of the different segmentation methods for one subject of the LPBA40 dataset.

Table 1 – Overall analysis against manual segmentation results for the *CC-12* subset. The best two values and all values better than auto-context CONSNet are emboldened. SSSMD = symmetric surface-to-surface mean distance.

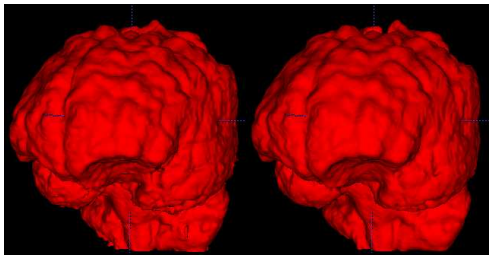
Methods	**Metrics**				
	Dice (%)	**Sensitivity** (%)	**Specificity** (%)	**Hausdorff** (mm)	**SSSMD** (mm)
ANTs	95.927 ± 0.009	94.510 ± 0.016	$\mathbf{99.705 \pm 0.001}$	$\mathbf{8.905 \pm 1.393}$	0.057 ± 0.015
BEAST	95.766 ± 0.012	93.838 ± 0.026	$\mathbf{99.757 \pm 0.001}$	9.907 ± 1.410	0.067 ± 0.029
BET	95.220 ± 0.009	98.261 ± 0.016	99.131 ± 0.002	12.169 ± 2.766	0.08 ± 0.024
HWA	91.657 ± 0.011	$\mathbf{99.930 \pm 0.001}$	97.830 ± 0.008	15.399 ± 1.799	0.179 ± 0.038
MBWSS	95.568 ± 0.015	92.784 ± 0.027	$\mathbf{99.848 \pm 0.004}$	28.228 ± 5.446	0.080 ± 0.031
OPTIBET	95.433 ± 0.007	96.133 ± 0.010	99.357 ± 0.003	10.304 ± 1.998	0.066 ± 0.013
ROBEX	95.611 ± 0.007	98.421 ± 0.007	99.130 ± 0.003	9.410 ± 1.610	0.063 ± 0.015
3D CNN	92.454 ± 0.032	88.770 ± 0.059	$\mathbf{99.648 \pm 0.001}$	21.244 ± 14.921	0.333 ± 0.349
STAPLE-12	96.797 ± 0.007	$\mathbf{98.976 \pm 0.006}$	99.382 ± 0.002	$\mathbf{8.327 \pm 1.665}$	0.038 ± 0.007
CONSNet	$\mathbf{97.183 \pm 0.005}$	98.919 ± 0.005	99.465 ± 0.002	9.713 ± 2.827	$\mathbf{0.037 \pm 0.007}$
auto-context CONSNet	$\mathbf{97.191 \pm 0.005}$	98.944 ± 0.005	99.465 ± 0.002	9.137 ± 2.374	$\mathbf{0.037 \pm 0.007}$



(a) Manual (b) STAPLE-OASIS $prob > 0.5$ (c) ANTs (d) BEAST (e) BET



(f) HWA (g) MBWSS (h) OPTIBET (i) ROBEX (j) 3D CNN



(k) CONSNet (l) auto-context CONSNet

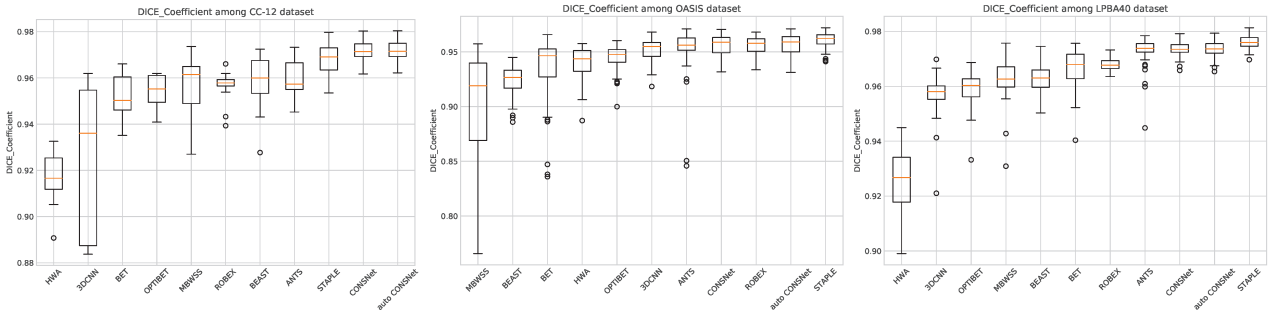
Figure 6 – Representative 3D reconstruction of the different segmentation methods for one subject of the OASIS dataset.

Table 2 – Overall analysis against manual segmentation results for the LPBA40 dataset. The best two values and all values better than auto-context CONSNet are emboldened. SSSMD = symmetric surface-to-surface mean distance.

Methods	**Metrics**				
	Dice (%)	**Sensitivity** (%)	**Specificity** (%)	**Hausdorff** (mm)	**SSSMD** (mm)
ANTs	97.259 ± 0.006	$\mathbf{98.981 \pm 0.004}$	99.179 ± 0.002	$\mathbf{9.394 \pm 3.876}$	$\mathbf{0.039 \pm 0.017}$
BEAST	96.306 ± 0.005	94.060 ± 0.012	$\mathbf{99.759 \pm 0.003}$	$\mathbf{9.447 \pm 3.724}$	0.058 ± 0.016
BET	96.625 ± 0.007	97.236 ± 0.014	99.276 ± 0.002	18.127 ± 6.379	0.079 ± 0.065
HWA	92.515 ± 0.012	$\mathbf{99.898 \pm 0.001}$	97.092 ± 0.006	16.110 ± 2.701	0.206 ± 0.055
MBWSS	96.239 ± 0.008	94.406 ± 0.013	$\mathbf{99.680 \pm 0.002}$	23.661 ± 6.283	0.075 ± 0.087
OPTIBET	95.874 ± 0.006	93.349 ± 0.011	$\mathbf{99.742 \pm 0.002}$	12.536 ± 2.838	0.064 ± 0.020
ROBEX	96.773 ± 0.002	96.491 ± 0.008	99.469 ± 0.002	12.472 ± 3.816	0.050 ± 0.006
3D CNN	95.696 ± 0.007	92.614 ± 0.015	$\mathbf{99.831 \pm 0.001}$	15.553 ± 5.062	0.070 ± 0.021
STAPLE-LPBA40	$\mathbf{97.585 \pm 0.002}$	$\mathbf{98.144 \pm 0.006}$	99.457 ± 0.002	$\mathbf{9.399 \pm 3.74}$	$\mathbf{0.033 \pm 0.005}$
CONSNet	97.353 ± 0.003	97.257 ± 0.007	$\mathbf{99.541 \pm 0.001}$	12.350 ± 3.721	$\mathbf{0.039 \pm 0.006}$
auto-context CONSNet	$\mathbf{97.356 \pm 0.003}$	97.330 ± 0.007	99.528 ± 0.001	11.991 ± 4.043	$\mathbf{0.039 \pm 0.007}$

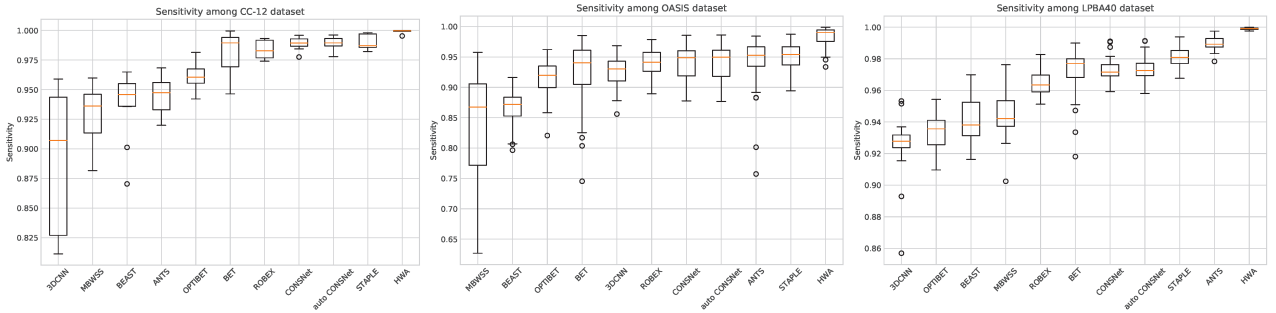
Table 3 – Overall analysis against manual segmentation results for the *OASIS* dataset. The best two values and all values better than auto-context CONSNet are emboldened. SSSMD = symmetric surface-to-surface mean distance.

Methods	**Metrics**				
	Dice (%)	**Sensitivity** (%)	**Specificity** (%)	**Hausdorff** (mm)	**SSSMD** (mm)
ANTs	95.307 ± 0.019	$\mathbf{94.391 \pm 0.036}$	98.732 ± 0.008	9.898 ± 4.350	0.114 ± 0.171
BEAST	92.468 ± 0.013	86.763 ± 0.025	$\mathbf{99.700 \pm 0.003}$	11.991 ± 1.905	0.167 ± 0.039
BET	93.503 ± 0.027	92.638 ± 0.048	98.101 ± 0.013	20.091 ± 6.768	0.227 ± 0.242
HWA	93.954 ± 0.014	$\mathbf{98.36 \pm 0.015}$	96.125 ± 0.016	14.062 ± 1.162	0.149 ± 0.055
MBWSS	90.241 ± 0.044	84.094 ± 0.079	$\mathbf{99.351 \pm 0.005}$	13.395 ± 8.086	0.249 ± 0.297
OPTIBET	94.456 ± 0.011	91.519 ± 0.027	$\mathbf{99.222 \pm 0.005}$	11.202 ± 1.714	0.110 ± 0.031
ROBEX	95.557 ± 0.008	93.954 ± 0.022	99.067 ± 0.005	$\mathbf{9.442 \pm 1.813}$	$\mathbf{0.083 \pm 0.025}$
3D CNN	95.237 ± 0.009	92.81 ± 0.023	$\mathbf{99.277 \pm 0.004}$	10.644 ± 2.642	0.095 ± 0.031
STAPLE-OASIS	$\mathbf{96.096 \pm 0.007}$	$\mathbf{95.188 \pm 0.02}$	98.983 ± 0.006	$\mathbf{8.553 \pm 1.602}$	$\mathbf{0.069 \pm 0.018}$
CONSNet	95.548 ± 0.010	93.98 ± 0.028	99.055 ± 0.006	10.228 ± 3.932	$\mathbf{0.083 \pm 0.028}$
auto-context CONSNet	$\mathbf{95.602 \pm 0.01}$	94.021 ± 0.028	99.078 ± 0.006	9.614 ± 3.658	$\mathbf{0.083 \pm 0.029}$



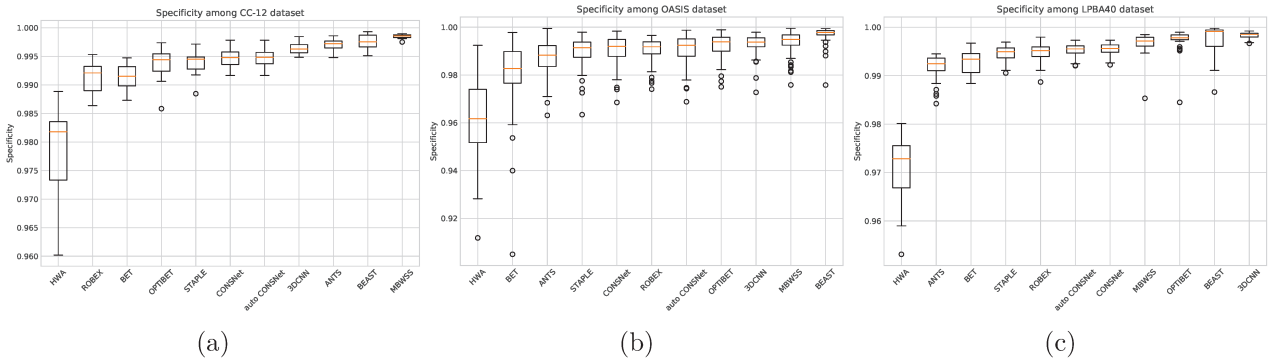
(a) (b) (c)

Figure 7 – Box plots of average Dice coefficient for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data.



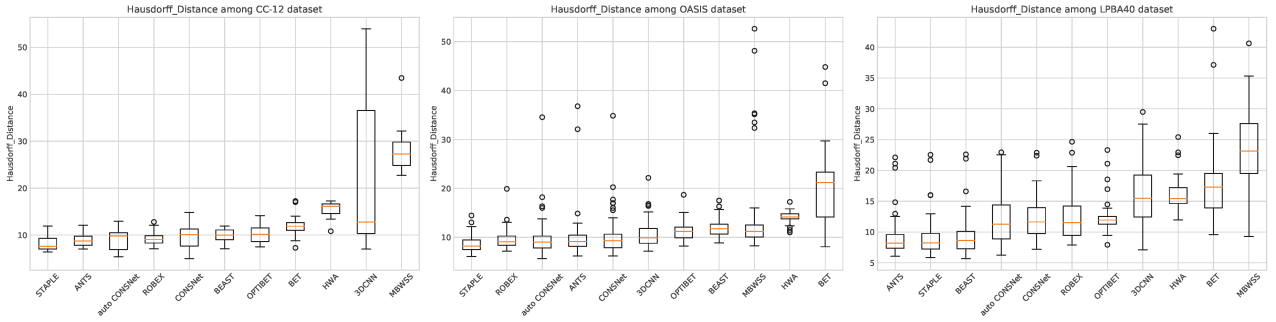
(a) (b) (c)

Figure 8 – Box plots of average sensitivity for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data.



(a) (b) (c)

Figure 9 – Box plots of average specificity for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data.



(a) (b) (c)

Figure 10 – Box plots of average Hausdorff distance for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data.

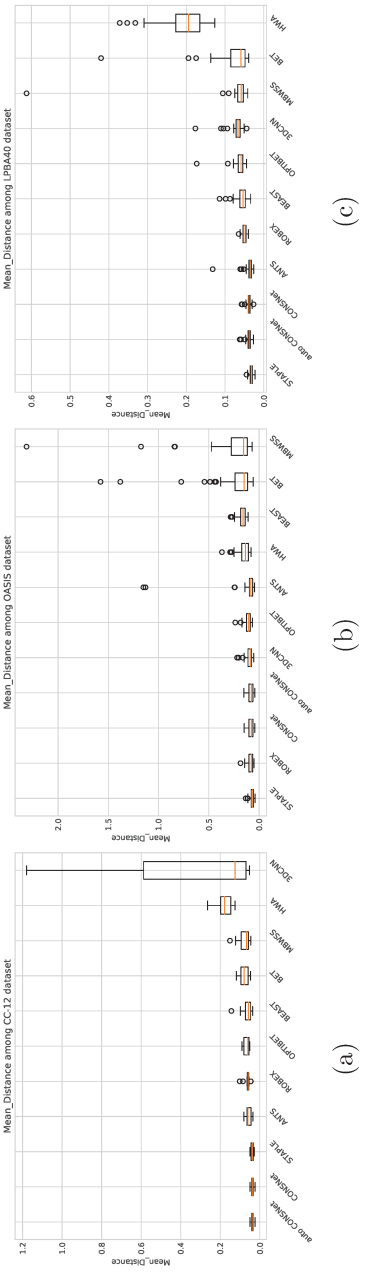


Figure 11 – Box plots of average symmetric surface-to-surface mean distance for each dataset: (a) *CC-12*, (b) OASIS, and (c) LPBA40. The boxes in the plots are sorted in the ascending order with respect to their mean value. BSE results were excluded due to poor results to allow for presentation of the data.

In general, auto-context CONSNet performed better than all other approaches (Dice coefficient), including CONSNet. Similar findings were observed over the other evaluation metrics, except for specificity. Auto-context CONSNet outperformed the 3D CNN method. The observed performance in the LPBA40 and OASIS datasets further suggest that this approach is generalizable.

Figures 12 to 14 present corresponding p-value heat-maps highlighting the statistical significance of each comparison *versus* the auto-context CONSNet results. Except in the OASIS dataset, where our CONSNet method ranked behind ROBEX, performance was good with the LPBA40 and OASIS datasets. The auto-context CONSNet model exhibited similar performance.

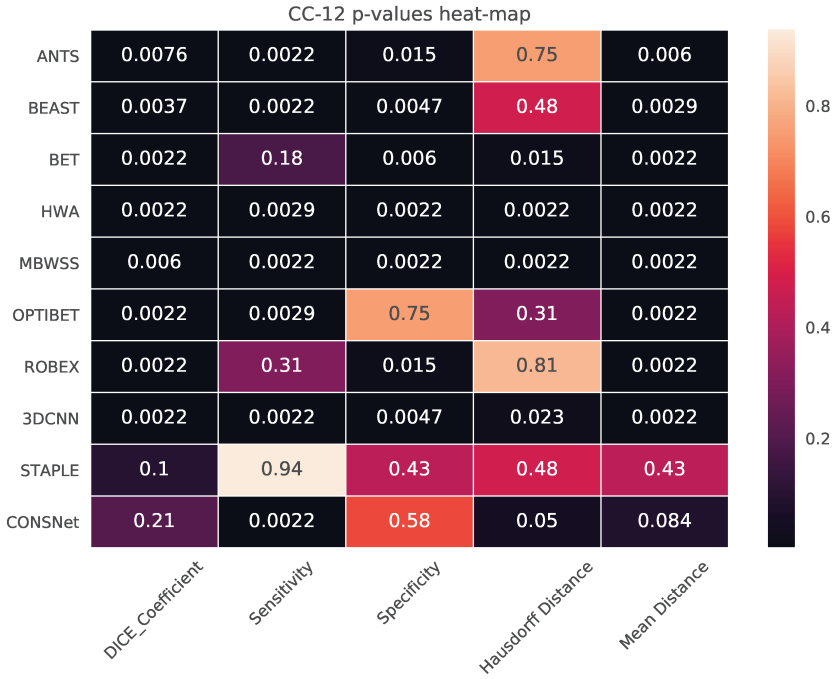


Figure 12 – Heat-map of the p-values calculated for the auto-context CONSNet across all evaluation metrics assessed in the *CC-12* subset. Darker cells highlight statistical significance (p-values < 0.05).

The heat maps (Figures 15 to 17) represent projections of the false positive (FP) and false negative (FN) errors. These sets of orthogonal projections highlight errors in the segmentations by method.

Processing times for all SS methods varied considerably (Table 4). The non DL-based approaches were CPU only based implementations. Both CPU and GPU versions of the DL-based approaches were implemented and both processing times are reported to allow comparison. Note that the processing time computed for the STAPLE method does not include the time required to execute the eight automated SS methods used as inputs; it only measured the processing time to formulate the consensus. The auto-context

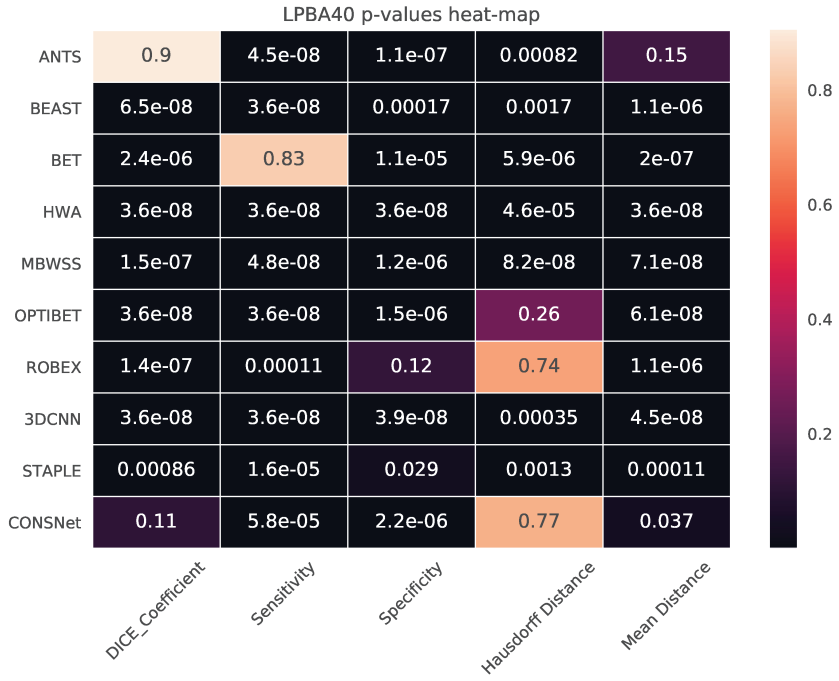
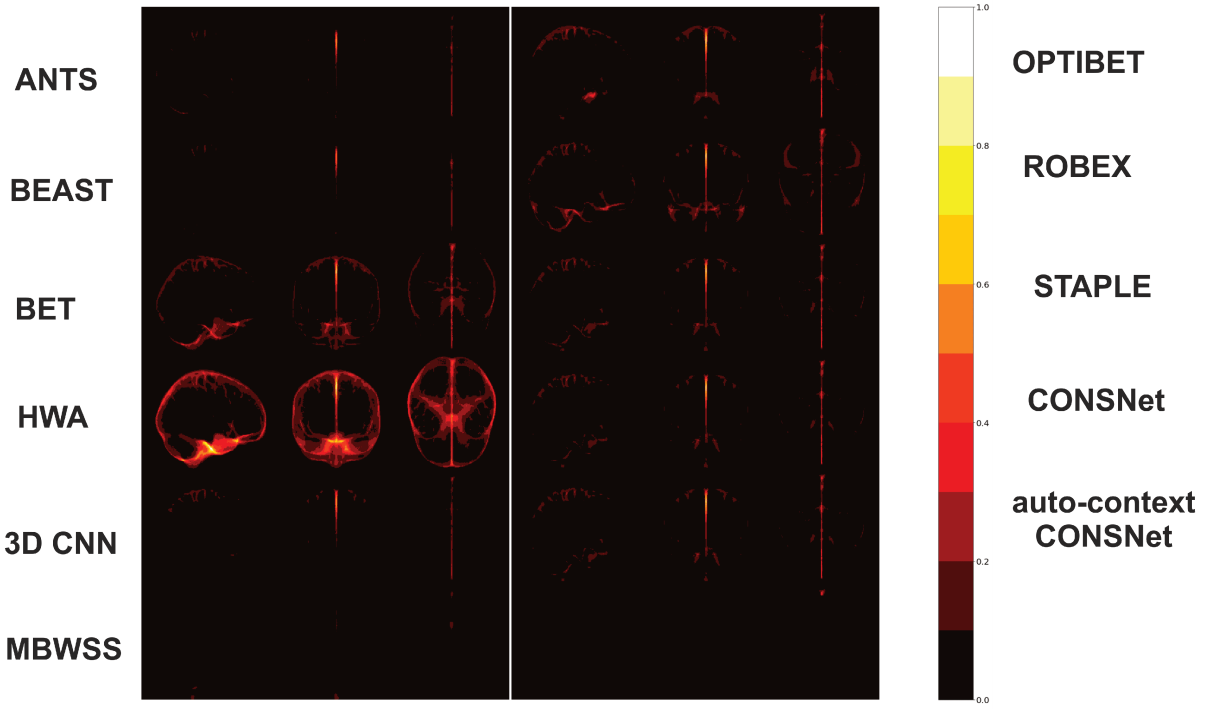
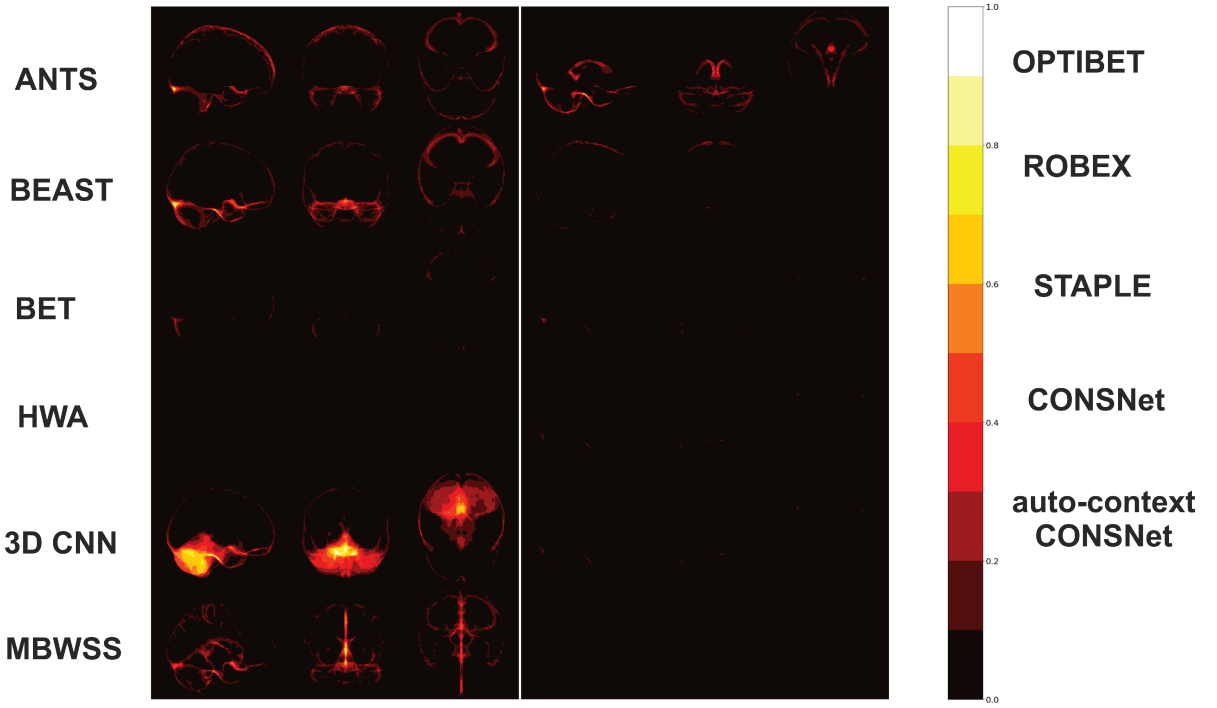


Figure 13 – Heat-map of the p-values calculated for the auto-context CONSNet across all evaluation metrics assessed in the LPBA40 dataset. Darker cells highlight statistical significance (p-values < 0.05)

Figure 14 – Heat-map of the p-values calculated for the auto-context CONSNet across all evaluation metrics assessed in the OASIS dataset. Darker cells highlight statistical significance (p-values < 0.05)

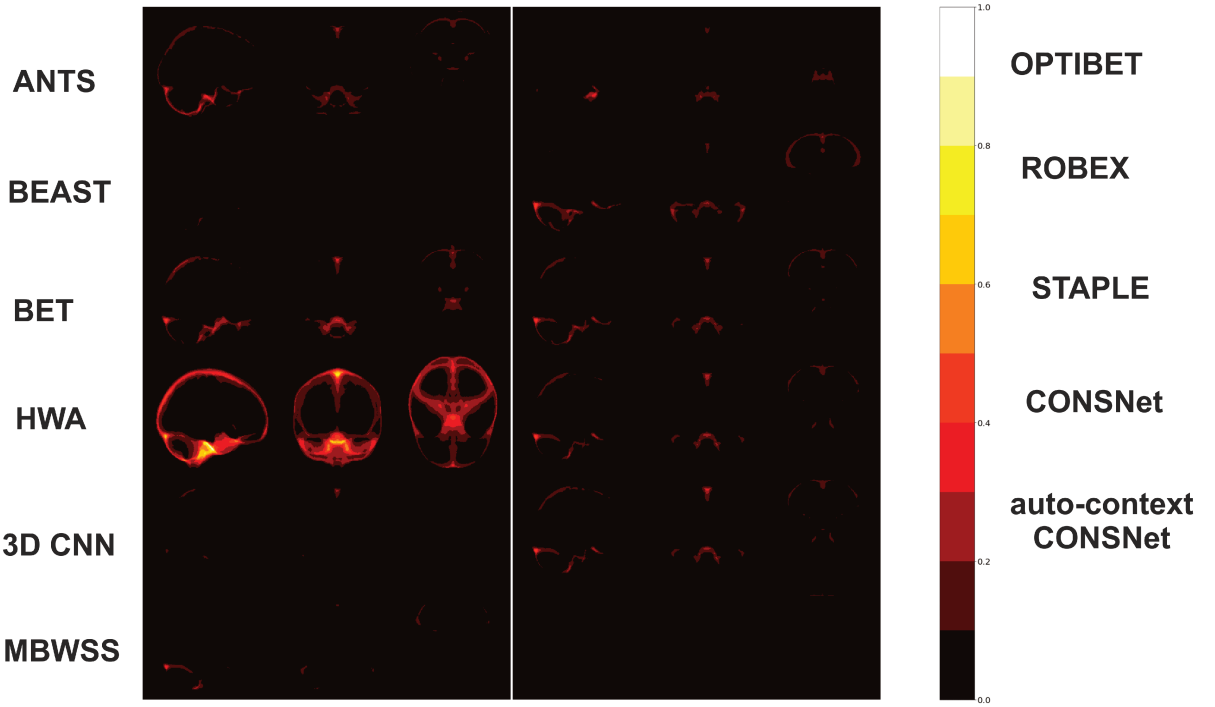


(a)

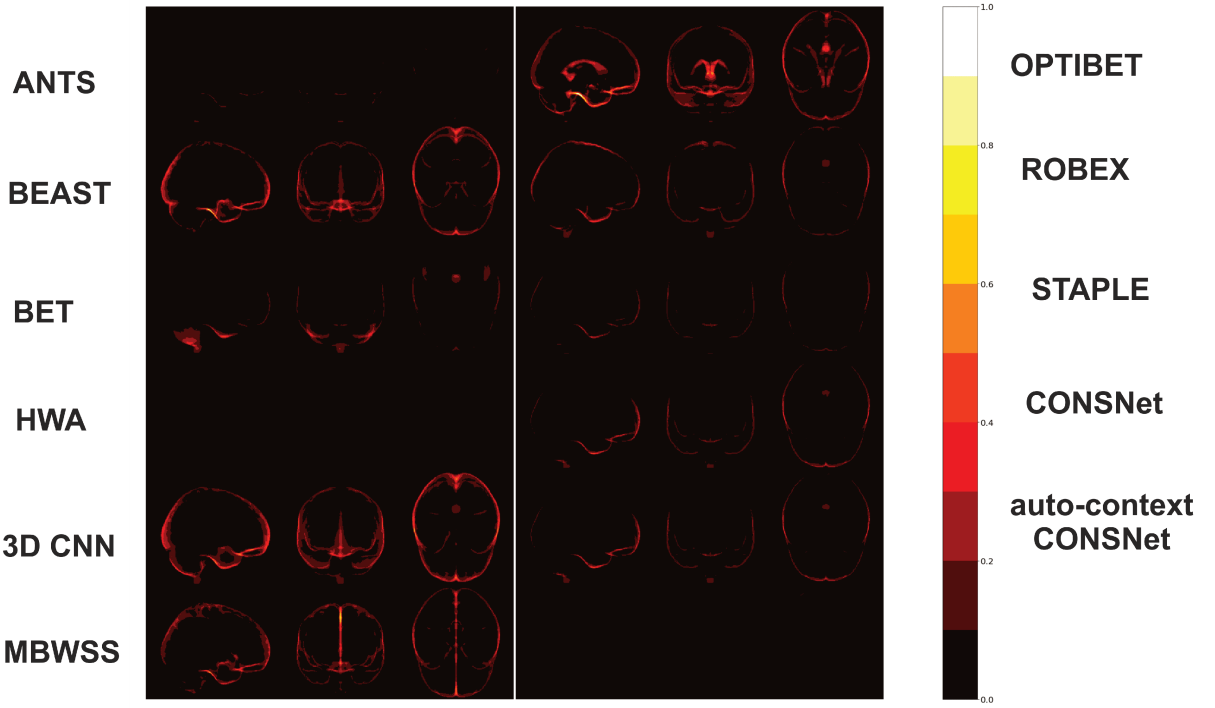


(b)

Figure 15 – Sagittal, coronal and axial heat map projections for the *CC-12* subset showing (a) false positive (FP) and (b) false negative (FN). The manual segmentations was used as the reference. Brighter voxels represents a high systematic number of FPs or FNs.

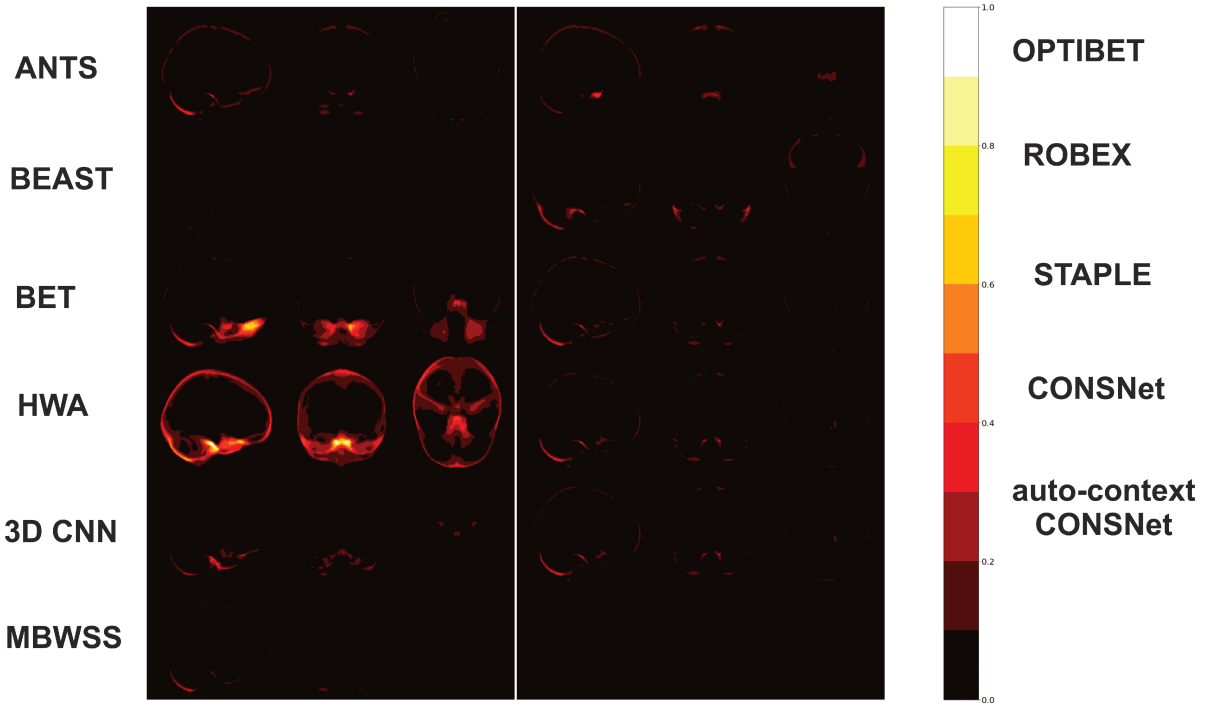


(a)

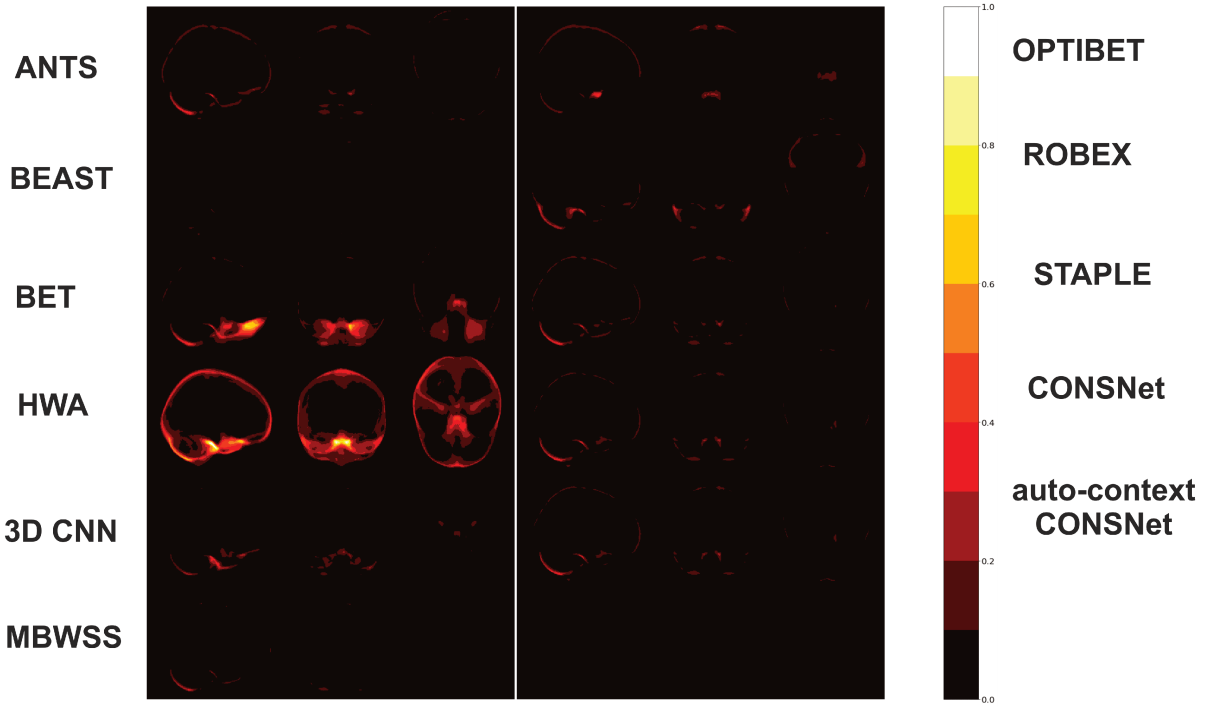


(b)

Figure 16 – Sagittal, coronal and axial heat map projections for the LPBA40 dataset showing (a) false positive (FP) and (b) false negative (FN). The manual segmentations was used as the reference. Brighter voxels represents a high systematic number of FPs or FNs.



(a)



(b)

Figure 17 – Sagittal, coronal and axial heat map projections for the OASIS dataset showing (a) false positive (FP) and (b) false negative (FN). The manual segmentations was used as the reference. Brighter voxels represents a high systematic number of FPs or FNs.

CONSNet processing time did not include the time to generate the probability maps from the CONSNet prediction stage. The auto-net is not included in these analysis since its source code is not publicly available.

Table 4 – Processing times for one image volume of each dataset (*CC-359*, OASIS, and LPAB40) for each skull-stripping method. For the CONSNet approaches, the number in front of the backslash represents the time computed on the CPU while the number after the backslash is the GPU time. * denotes results for the processing time for the STAPLE consensus-forming step only or the auto-context CONSNet step only (see text).

Processing time (seconds)			
Method	**Datasets**		
	CC-12	**OASIS**	**LPBA40**
ANTs	1378	1025	1135
BEAST	1128	944	905
BET	9	5	7
BSE	2	1	1
HWA	846	248	281
MBWSS	135	66	79
OPTIBET	773	579	679
ROBEX	60	53	57
3D CNN	196	121	123
STAPLE (12,OASIS,LPBA40)	160*	54*	36*
CONSNet	516/25	214/18	301/20
auto-context CONSNet	155*/11*	75*/8*	105*/10*

2.3.2 Experiment 2

Tables 5 and 6 summarize the results of Experiment 2. Evaluation metrics are shown for the two-fold cross-validation experiment using the LPBA40 and OASIS datasets for training. Also summarized are the evaluation metrics published for the 3D CNN (KLEESIEK *et al.*, 2016), and for auto-net and U-Net (SALEHI *et al.*, 2017) networks using the LPBA40 data (Table 5) and the OASIS (Table 6) datasets. Comparison is also made validating on the *CC-12* subset. Our results compared favourably to the result published using the 3D CNN (KLEESIEK *et al.*, 2016), auto-net and U-Net (SALEHI *et al.*, 2017) methods. CONSNet was always best or second best across the five evaluation metrics.

2.4 Discussion

2.4.1 Experiment 1

Our approaches had excellent results with respect to the Dice coefficient which is the first metric taken by the raters to evaluate an optimal segmentation. CONSNet also

Table 5 – Two-fold cross-validation using the LPBA40 dataset. The best score using the LPBA40 dataset for each metric is emboldened. Values for 3D CNN (KLEESIEK *et al.*, 2016), and for auto-net and U-Net (SALEHI *et al.*, 2017) are from literature. SSSMD = symmetric surface-to-surface mean distance.

Datasets	Methods	Metrics				
		Dice (%)	**Sensitivity** (%)	**Specificity** (%)	**Hausdorff** (mm)	**SSSMD** (mm)
CC-12	**CONSNet**	89.63 ± 0.076	85.11 ± 0.129	99.6 ± 0.002	16.67 ± 3.915	0.24 ± 0.217
LPBA40	**CONSNet**	$\mathbf{98.47 \pm 0.002}$	$\mathbf{98.55 \pm 0.005}$	$\mathbf{99.71 \pm 0.001}$	$\mathbf{10.05 \pm 5.087}$	$\mathbf{0.02 \pm 0.003}$
	Auto-net (SALEHI *et al.*, 2017)	97.73 ± 0.003	98.31 ± 0.006	99.48 ± 0.001	––	––
	U-Net (SALEHI *et al.*, 2017)	96.79 ± 0.004	97.22 ± 0.016	99.34 ± 0.002	––	––
	3D CNN (KLEESIEK *et al.*, 2016)	96.96 ± 0.01	97.46 ± 0.01	99.41 ± 0.003	––	––
OASIS	**CONSNet**	92.55 ± 0.03	89.11 ± 0.059	98.86 ± 0.007	13.09 ± 4.483	0.15 ± 0.075

Table 6 – Two-fold cross-validation using the OASIS dataset. The best score using the OASIS dataset for each metric is emboldened. Values for 3D CNN (KLEESIEK *et al.*, 2016), and for auto-net and U-Net (SALEHI *et al.*, 2017) are from literature. SSSMD = symmetric surface-to-surface mean distance.

Datasets	Methods	Metrics				
		Dice (%)	**Sensitivity** (%)	**Specificity** (%)	**Hausdorff** (mm)	**SSSMD** (mm)
CC-12	**CONSNet**	92.22 ± 0.022	94.17 ± 0.058	98.92 ± 0.004	18.55 ± 13.443	0.17 ± 0.087
LPBA40	**CONSNet**	92.31 ± 0.046	90.78 ± 0.08	99.0 ± 0.003	17.8 ± 8.306	0.18 ± 0.135
OASIS	**CONSNet**	97.14 ± 0.005	97.45 ± 0.013	$\mathbf{98.88 \pm 0.006}$	$\mathbf{6.9 \pm 1.549}$	$\mathbf{0.04 \pm 0.013}$
	Auto-net (SALEHI *et al.*, 2017)	$\mathbf{97.62 \pm 0.01}$	$\mathbf{98.66 \pm 0.01}$	98.77 ± 0.01	––	––
	U-Net (SALEHI *et al.*, 2017)	96.22 ± 0.006	97.29 ± 0.01	98.27 ± 0.007	––	––
	3D CNN (KLEESIEK *et al.*, 2016)	95.02 ± 0.01	92.40 ± 0.03	99.28 ± 0.004	––	––

had excellent sensitivity (*i.e.*, keeping brain tissue in the SS mask). In fact, all methods performed equivalently with the exception of HWA, which usually had a high sensitivity but poor overlap (Dice coefficient) and specificity due to the challenges of atlas-based registration. With respect to specificity, all methods performed acceptably, except for HWA (Figure 9). Also, our CONSNet methods had very few outliers in the Hausdorff and symmetric surface-to-surface mean distances (Figures 10 to 11). Using the auto-context CONSNet as a reference, some metric differences were found to be statistically significant (*i.e.*, p-value < 0.05) compared to most other methods (Figures 12 to 14). STAPLE and ANTs, had fewest statistically significant performance differences.

The projection heat maps showed interesting average differences between the methods. Figure 15a, for example, shows that methods with high specificity, such as ANTs and BEAST were not able to properly segment the fissure between the left and right brain hemispheres in the *CC-12* data subset. Only MBWSS was capable of correctly segmenting the brain fissure. In the STAPLE consensus algorithm, MBWSS is one out of eight methods and its influence is diluted by the other algorithms that did not correctly segment the brain fissure, thus affecting the silver standard mask used to train CONSNet. Our approach was more influenced by the other methods and, as a result, did not segment the brain fissure correctly. Both OASIS and LPAB40 have annotated masks with smoothed brain fissures. From Figures 16a and 17a it is possible to see that we had very few FP

errors even along the brain border region.

The 3D CNN method had the highest number of FN errors than the other methods in the *CC-12* data subset. Figures 15b, 16b and 17b show that we had few FN in all datasets, mostly in the OASIS dataset. In this heat maps analysis, STAPLE had similar results to our method, except in the OASIS and LPBA40 datasets.

The guidelines for manual annotation varies between expert raters, influencing the performance evaluation of the predictions for different datasets (ESKILDSEN *et al.*, 2012). The *CC-12* subset had a manual annotation which correctly segmented the gyral and sulcal regions as well as the brain fissure. The LPBA40 dataset had a comparatively smoother surface delineation. The OASIS dataset had annotated brain mask data that was generated using automatic segmentation and then manually revised by an expert. Our silver standard masks were the agreement among different traditional algorithmic approaches, overcoming the possible super-specialization in the CNN training when using only one procedure for generating annotated data. As a result, we have an optimal performance across datasets when CONSNet and auto-CONSNet were trained with silver standard masks and validated with LPBA40 and OASIS datasets.

It was observed that the processing time of the most robust individual methods (ANTs, BEAST) was $\approx$ 20 minutes for larger image volumes like in the *CC-12* data subset (Table 4). ROBEX which profits from a parallel implementation, was the fastest, taking around a minute for the same image volume. Our CPU-based implementation of the auto-context CONSNet prediction took over two minutes without considering the time to generate the probability maps from the parallel CNNs predictions. If we consider this additional time, the auto-context CONSNet prediction required $\approx$ 11 minutes (671 seconds, which equals $3\times$ the time to generate the probability maps plus the time of the auto-context CONSNet prediction itself, or $516+155$ seconds). This aggregate time is still approximately half that required by ANTs or BEAST. CONSNet prediction times were significantly reduced in the GPU implementation, our method requiring $\approx$ 25 seconds.

STAPLE as a method outperformed our method in almost all of the computed metrics in the OASIS and LPBA40 dataset (Tables 2 and 3) with statistical difference (Figures 13, and 14). These results suggest that our CONSNet results might be limited by the STAPLE results. Nonetheless, using STAPLE as a method is very expensive and time-consuming because for every new image volume the automatic methods need to be ran again to generate a new STAPLE brain mask. In other words, maximum performance is determined by the standard used for annotation (in this work, that is, the STAPLE-derived, silver standard consensus). As an additional note, the quality of the manually segmented data in both OASIS and LPBA40 datasets are inferior than the *CC-12*. This could also limit our observed performance. More investigation of these limitations, therefore, should be undertaken.

2.4.2 Experiment 2

Experiment 2 consisted of a two-fold cross validation experiment using first the LPBA40 and then the OASIS dataset for training. We explored only the CONSNet method because the approach had similar results to the auto-context CONSNet but was faster (see 4). CONSNet also had comparable performance to the auto-net and outperformed the U-Net and the 3D CNN methods (Tables 5 and 6).

CONSNet performed better than all DL-based methods in the LPBA40 dataset across all metrics. In the OASIS dataset, we ranked second in Dice coefficient and in sensitivity; but were ranked first regarding specificity. The U-Net pipeline from (SALEHI *et al.*, 2017) work does not have an auto-context CNN. Therefore, a fairer comparison between the CNN architectures would be to compare against U-Net. CONSNet outperformed U-Net by 1% and 1.7% margin in the OASIS and LPBA40 datasets, respectively. Our CONSNet architecture was very robust, having Gaussian noise, dropout, and batch normalization as regularizers. These steps lead to a substantial improvement against U-Net (SALEHI *et al.*, 2017) and comparable results to auto-net (SALEHI *et al.*, 2017).

Two-fold cross validation analysis cannot demonstrate generalization. When the two-fold-derived models were applied to datasets with images that were not included in the training step, the best results with respect to Dice coefficient, for instance, ranged from 89% to 92%. This result occurred because CNNs are data-driven approaches and during their training stage they only learn to reproduce the annotation (rules) provided by the training data. Therefore they are biased to the provided manual annotation. As a consequence, since LPBA40 and OASIS are not robust datasets (*i.e.*, only one vendor and scanner, and have poor annotation), good generalization was not achieved.

2.4.3 General Discussion

Overall, CONSNet is the most appropriate choice between our two pipelines approaches (CONSNet and auto-context CONSNet). This model has low cost (fast) and achieved similar performance compared to auto-context CONSNet in Experiment 1. It also achieved similar results to auto-net in Experiment 2 but with lower cost. These results show that by using silver standard data as annotated input for training results in better generalization (Experiment 1) compared to the two-fold cross-validation with gold standard, manual masks (Experiment 2). As discussed, silver standard data are generated through agreement among different automated processes (each with distinct guidelines). This approach can reduce super-specialization effects in training and minimize dominance of one expert/process/guideline on the CNN training. Silver standard data can also minimize the impact of inter-/intra- rater variability in the annotation step. Additionally, we generated silver standard masks from unlabeled data, greatly enlarging our input annotated data: by adopting a patch-wise training scheme with more than 100,000 input

data.

2.5 Conclusions

In this chapter, we have proposed a robust CNN for brain MR imaging SS, fully trained using silver standards masks. CONSNet and auto-context CONSNet are comparable to state-of-the-art automatic approaches, faster than the most robust non DL-based methods (even CPU-based implementations), and have optimal generalization. The usage of silver standard masks, also, provides a low cost solution for generating annotated input data to augment the volume of training data. Silver standards also reduce inter-/intra-rater variability and decrease overfitting of data-driven approaches because the annotations are generated through consensus agreement. With these results we want to leverage the usage of silver standard brain masks for large-scale studies in medical image processing. Moreover, we also want to extend CONSNet to a 3D architecture. We recognize the paucity of expertly annotated data in other medical image processing tasks and, in the future, want to extend our contribution to other applications.

3 Deep-learning-based Tractography for Surgical Planning in Epilepsy treatment

3.1 Motivation

Epilepsy is a brain disorder characterized predominantly by recurrent and unpredictable interruptions of normal brain function, called epileptic seizures (FISHER *et al.*, 2005). Epilepsy affects 50 million [1] people worldwide. Between 20 and 40 % of focal epilepsy patients are refractory, i.e., do not respond to treatment with antiepileptic medication (KWAN; BRODIE, 2004). An effective treatment for these patients is surgical resection of the epileptogenic zone (EZ) (WIEBE *et al.*, 2001; PLATT; SPERLING, 2002), which is defined as the minimal cortex area that must be resected to produce seizure-freedom (LÜDERS *et al.*, 2006). Surgical resection of the hippocampus is performed to cure refractory temporal lobe epilepsy (WINSTON *et al.*, 2012; WINSTON *et al.*, 2011; KWAN; BRODIE, 2004), consisting of either a specific resection of the hippocampus and amygdala or a complete anterior temporal lobe resection. In this surgery, there is a 10% risk of a significant visual field deficit (GOONERATNE *et al.*, 2017) because of damage to the optic radiation, the white matter (WM) fibres of the visual system (WINSTON *et al.*, 2011).

Tractography is a non-invasive method for visualization of the white matter fiber bundles or tracts done using information in diffusion magnetic resonance images (dMRI) (YAMADA *et al.*, 2009). Tractography identifies white matter pathways based on the assumption of predominant water diffusion along the fibers rather than perpendicular to them (LEMKADDEM *et al.*, 2014). The primary clinical application of tractography is preoperative planning (YAMADA *et al.*, 2009), for example the delineation provided by tractography of the optic radiation, which cannot be performed using conventional MR imaging sequences, is used to determine the distance between the temporal pole and Meyer's loop (WINSTON *et al.*, 2012; WINSTON *et al.*, 2011), consequently reducing the risk of damage at the visual field. The goal of presurgical tractography is to identify the position of eloquent pathways, such as the motor, sensory, and language tracts (BERMAN, 2009) to plan the surgery aiming to avoid damaging these bundles.

3.1.1 Previous Works in Tractography

Although tractography has demonstrated huge benefits in both neuroscience and clinical applications, there are outstanding challenges in how to get the most out of the

[1] http://www.who.int/news-room/fact-sheets/detail/epilepsy

acquired dMRI data. Usually, tractography is performed in two main steps: the first step is to obtain a voxel-wise estimate of local fiber orientations and the second step is to track fibers across voxels (ZHAN *et al.*, 2015). There are a plethora of methods to convert acquired dMRI data into local fiber orientations which are then used in tractography, for instance, using diffusion tensor (BASSER *et al.*, 1994), multi-tensor (MALCOLM *et al.*, 2010), spherical deconvolution (TOURNIER *et al.*, 2004), or multi-compartment models (ASSAF *et al.*, 2008). Three main approaches exist to perform tractography which are deterministic, probabilistic, and global (LEMKADDEM *et al.*, 2014; MORI *et al.*, 1999; YENDIKI *et al.*, 2011; TOURNIER *et al.*, 2012).

In general, deterministic tractography algorithms follow the peak fiber orientation at discrete locations in small, discrete steps (ALEXANDER, 2010). The other approaches use multiple start fibers, and then find multiple directions for a given voxel. Initially, tractography is performed by starting from one or more "seed" locations propagating the trajectories according to some constraints until the tracts are terminated (ALEXANDER, 2010). Probabilistic tractography relies on a probability density function (PDF) for fiber orientation (PARKER, 2010). In this approach, a probability is assigned to the reconstructed pathways by considering multiple pathways emanating from the same seed (JEURISSEN *et al.*, 2011). Global methods try to reconstruct all the fibers simultaneously by finding the configuration that describes best the measured data, this necessarily involves the solution of the forward problem of predicting the measured signal given a fiber configuration (REISERT *et al.*, 2011). For these methods, anatomical prior constraints are often used.

The conventional tractography methods have several unresolved challenges: how to distinguish between different complex fiber configurations within a voxel, where axons can cross, kiss, bend, or fan out; premature termination of tracts based on predefined stopping criteria; or the trade-off between specificity and sensitivity of different approaches (in either local orientations or tractography) (NEHER *et al.*, 2015; NEHER *et al.*, 2017). These issues in data processing for tractography are intricately tied to data acquisition parameters, such as the minimum number of diffusion-weighting gradients and the chosen b-values (BERMAN, 2009; LEMKADDEM *et al.*, 2014; NEHER *et al.*, 2015; NEHER *et al.*, 2017). Moreover, the selection of specific fiber bundles, like the optic radiation, is typically done as a post-processing step independent of signal modeling, using expert manual input which can be very time consuming and has significant variability between manual raters (YAMADA *et al.*, 2009; NOWELL *et al.*, 2015).

Recently, machine-learning (NEHER *et al.*, 2017) and deep-learning-based (POULIN *et al.*, 2017) tractography approaches have been proposed, which use the measured signal directly and rely on machine learning advantages, such as neighborhood information to disentangle asymmetric fiber patterns. The lack of general restrictions on the type of im-

age acquisition forgo the necessity of ad hoc constraints of signal modeling (NEHER *et al.*, 2017). Moreover, deep-learning-based methods have the advantage to be fast at the prediction stage (LECUN *et al.*, 2015). However, none of the aforementioned approaches have used convolutional neural networks (CNNs).

3.1.2 Our Approach

We present in this chapter a deep-learning-based approach to perform tractography for surgical planning in epilepsy treatment. We structured the problem in a deep-learning-regression pipeline to predict the fibers bundles, using a three-stage approach: (1) use an auto-encoder to learn a high-level representation in a latent space for the coordinates in each fiber; (2) regressing the input images to map the high-level representation of tract by feeding them into a 3D 18-convolutional-layer residual network (ResNet18); (3) fine-tuning the stack model (ResNet18 + decoding layers from the auto-encoder) to regress the tracts. We validated our approach using 10 subjects in a five-fold cross-validation by doing two experiments: (a) training and testing in a subject and (b) training in data from all subjects and testing the model for each one. We used Dice coefficient as the evaluation metric to compute the overlap area between the voxels that both prediction and ground truth fiber bundles were passing through in the raw data.

The main contributions in this work can be summarized as follows:

- To the best of our knowledge, we are the first ones to investigate the use of CNNs for tractography.
- Our method uses a three stages approach with an auto-encoder to obtain a high-level representation of the fiber bundles, making the final regression more robust.
- Our analysis serves as a baseline for deep-learning-based regression for tractography in epilepsy surgery.

3.2 Materials and Methods

3.2.1 Dataset

Our dataset is comprised of 10 subjects. Each dMRI is acquired in multi-shell format with the following characteristics: $2mm$ isotropic resolution, with 115 gradient directions: 11, 8, 32, and 64 at b-values: 0, 300, 700, and $2500 s/mm^2$, single $b = 0$ image with reverse phase-encoding to correct for susceptibility-induced distortions (ANDERSSON *et al.*, 2003). Each patient has annotation of the uncinate fasciculus (UF) tract which is composed of 5000 fibers (Figure 18). Each fiber has a set of points in which the coordinates

represent the 3D position in *mm*. Fiber tracts were reconstructed using the probabilistic tractography method from MRTrix3 (TOURNIER *et al.*, 2012) using integration over fiber orientation distributions (TOURNIER *et al.*, 2010) and anatomically-constrained tractography, then, they were manually revised by two experts, following an established criteria (WAKANA *et al.*, 2007).

Figure 18 – 3D reconstruction of the UF tract.

3.2.2 Convolutional Neural Network Architectures and Implementation

We structured the tractography problem in a deep-learning-based regression and conducted the experiments using two networks. The first network is an autoenconder with 1 hidden layer comprised of 1024 neurons, l_2 weight regularization of 10^{-4}, l_1 activity regularization of 10^{-5}, scaled exponential linear units (SeLU) (KLAMBAUER *et al.*, 2017) activation for the hidden layer, and linear activation for the output layer. The second one is a 3D 18-convolutional-layer residual network (ResNet18) (HE *et al.*, 2016) with ReLU activations in all layers except for the last layer which contains a linear activation. Both were implemented using Keras with Tensorflow (ABADI *et al.*, 2016) as a backend, and the ResNet18 was based on a publicly available code [2]. In this CNN, we modified the stride to be $1 \times 1 \times 1$ in both first convolution and max-pooling layer. Figure 19 depicts the adopted ResNet18 architecture.

Nadam (DOZAT, 2016) was used as the optimizer with an initial learning rate of 0.001 and a step decay of after every five epochs at the training stage for ResNet18. For the

[2] https://github.com/JihongJu/keras-resnet3d

autoencoder training, we adopted Adam (KINGMA; BA, 2014) optimizer with a learning rate of 0.0001 and no weight decay. The Huber loss presented in Equation 3.1 was used as the cost function in both networks. We chose this function because for regression tasks it is less sensitive to outliers in the training data than the mean squared error (ZINKEVICH *et al.*, 2010).

$$h(x) = \begin{cases} 0.5x^2, & \text{if } |x| \leq d \\ 0.5x^2 + (|x| - d), & \text{otherwise} \end{cases} \tag{3.1}$$

$$\text{where } d = 1.0,\ x = y_{true} - y_{pred}$$

3.2.3 Proposed Tractography Pipeline

To perform the regression, we based our approach on learning a high-level latent representation used in (KATIRCIOGLU *et al.*, 2018) to predict 3D human pose estimation. In this work, the authors show that combining a traditional CNN for supervised learning with autoencoders results in an improved performance since it learns to properly encode dependencies between joint locations.

First, they learned high-level representation embeddings using autoencoders. Secondly, a regression network is trained to map an input to these embeddings. As a final step, they stacked the decoding layers of the trained autoencoder on top of the CNN and fine-tuned the whole architecture to predict the joint locations. Based on their approach, we built our pipeline consisting of the same three stages shown in Figure 20. The three stages for our proposed tractography are: (a) learn high-level models on a latent space to represent the tracts with an autoencoder, in our case an embedding with size of 1024 (experimentally set); (b) regressing the input images to map a high-level representation of tract; (c) fine-tuning the stack model (ResNet18 + decoding layers) to regress the tracts.

3.2.3.1 Pre-processing

White matter tracts can vary in size due to patients' head size making the CNN training stage challenging when using data from different subjects potentially leading to wrong inferences (i.e. inferring a streamline in grey matter or lack of reproducibility) (MAIER-HEIN *et al.*, 2017; ESSAYED *et al.*, 2017). Another issue is using raw data as an input to feed a CNN. Ideally, raw data has more information than a processed data (fit into fiber orientation model) as a model signal orientation necessarily is constrained by the signal model. However, dMRI is noisy and raw data values may restrict the method by the used MRI acquisition of the training dataset, not allowing slight variations in the acquisition setup, such as patient orientation with the scanner, without a complete retraining of the method (WASSERTHAL *et al.*, 2018).

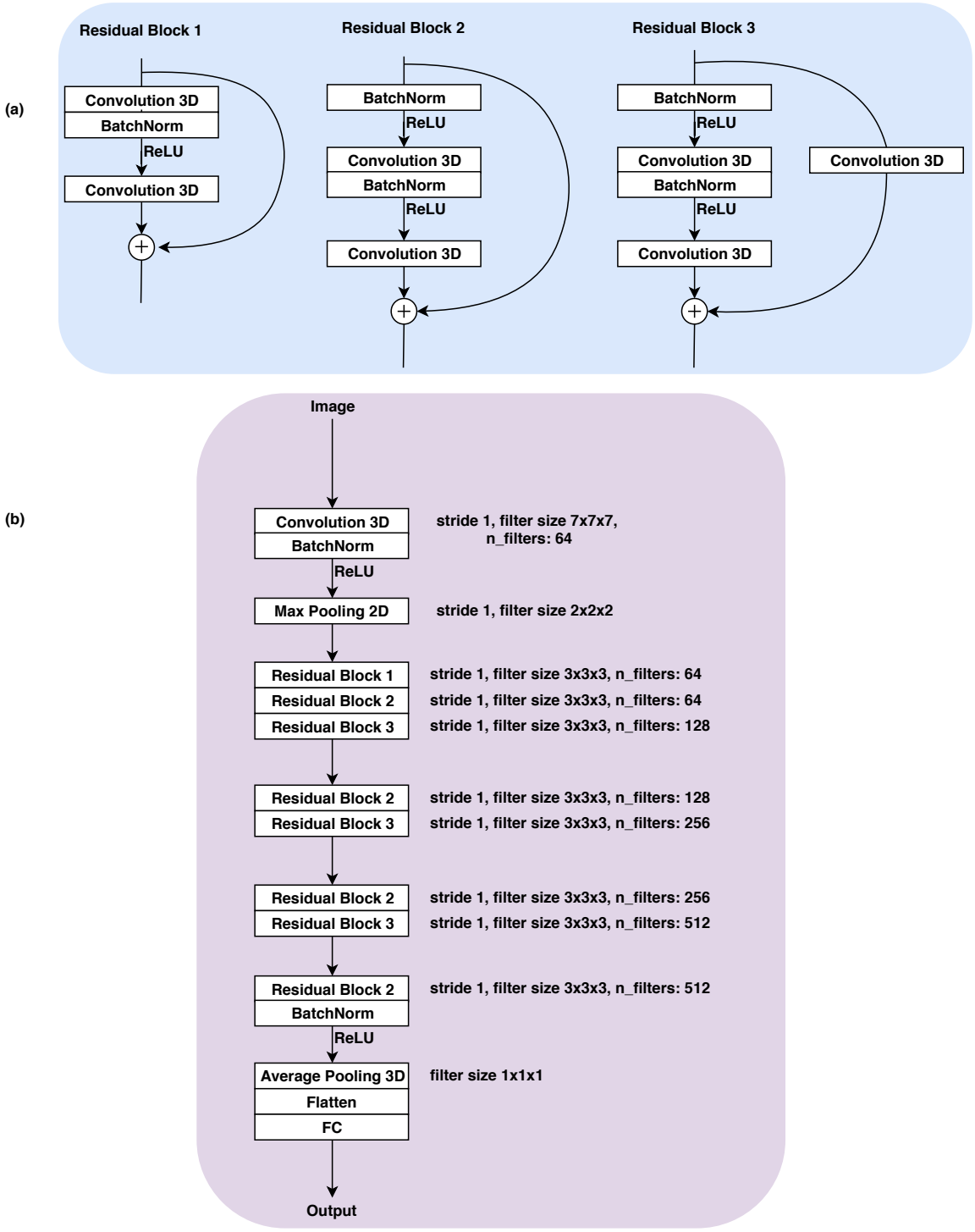


Figure 19 – 18-layer 3D ResNet implemented. (a) Presents the residual blocks implemented in the network. (b) Depicts the full architecture adopted where FC stands for fully connected layer.

To overcome the above issues, we used as input three principal fiber directions per voxel as input, resulting in 9 features per voxel like in (WASSERTHAL *et al.*, 2018). First, we fit our raw data into a constrained spherical deconvolution (CSD) model using

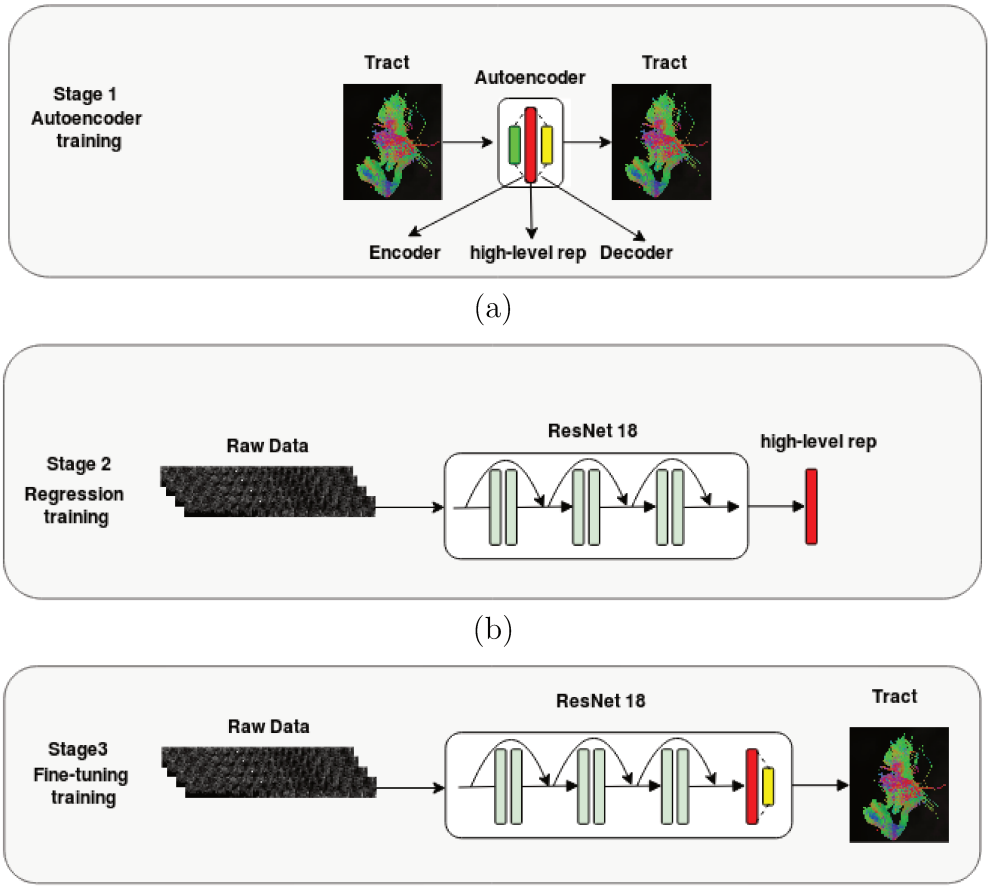


Figure 20 – Proposed tractography pipeline. (a) Autoencoder , (b) regression training to map the high-level embeddings representation, and (c) fine-tuning the whole network (ResNet18 + trained decoding layers)

MRtrix (TOURNIER *et al.*, 2012). Then, we extracted the three largest peaks from the CSD model. Moreover, to achieve reproducibility and leverage training data from different subjects, we register all the subjects data after peak extraction using also MRtrix through a symmetric diffeomorphic image registration proposed by (AVANTS *et al.*, 2008). This registration generates an unbiased group-average template from a series of images using an iterative averaging approach, symmetrically aligning them to a 'midway' space. Next, we resampled the fibers from each annotated tract to 120 coordinates, and then we applied the generated warp transformations.

3.2.3.2 CNN Regression

We aim at directly regressing from an input dMRI I to a 3D streamline S. Consequently, the problem consists of training a set of N images $I = \{I_1, ..., I_N\}$ and the associated ground truth fibers $S = \{s_1, ..., s_N\}$, $s_i \in \Re^{3P}$ consisting of coordinates re-

ferring to P points, which in our case is 120. Therefore, our approach aims to regress an input 3D data to a correspondent fiber. To reach that, our CNN regression was performed by feeding the networks with 3D patches. 3D patches were extracted in fixed size of $48 \times 48 \times 48$. They represent the white matter voxels that each fiber from the tract is passing through. The patches generation is done based on segmented mask from the tract. These segmented masks were generated by a nearest-neighbor interpolation.

3.2.4 Experimental Methodology

Our experiments were focused on the left UF tract because this bundle is a well-defined bundle in terms of end-points and despite the curved geometry has high reproducibility between raters. These features make the UF a good candidate to validate tractography performance.

We conducted two five-fold cross-validation experiments to evaluate the performance of the proposed method. In Experiment 1, we performed our CNN regression in one subject for training and testing, and the procedure was repeated for all ten patients. In this case, all 5000 fibers from the left UF tract were used. In Experiment 2, we were interested in evaluating the performance of our tractography method per patient by adding data from other subjects. In this case, we randomly extracted 700 fibers from each of the ten subjects and evaluated the trained model against each patient.

3.2.5 Evaluation Metrics

Using the Dice coefficient, we quantified the overlap between our predictions and the annotated tracts. This overlap is measured by retrieving brain regions per voxel of the dMRI that the fibers are passing though. We map the tracts to weighted high-resolution image, using MRtrix, and thresholded this image to output a binary mask when the probability weight was greater or equal to 0.5. The Dice coefficient was then calculated between these masks.

The statistical analysis to assess differences in the evaluation metrics was done using a t-test. A p-value less than 0.05 was used to confirm statistically significant differences.

3.3 Results

Table 7 summarize the averaged results for the five-fold cross validation in the Experiment 1 and 2. Figure 21 depicts a bar plot for a better comparative visualization of the Dice among patients in both experiments. In general, our approach did not achieve an optimal performance in both experiments. However, combining data from different

subjects in Experiment 2 resulted in a slight performance improvement compared to Experiment 1. Except for subject **6** and **7**, all patients had an enhanced overlap (i.e. improvement of Dice coefficient) for Experiment 2.

Figures 22 and 23 show a 3D reconstruction of the ground truth annotation and predicted tracts of the best and worst subject performances in Experiment 1 and 2 based on the scores of Table 7. The predicted tracts followed the geometry of the ground truth, however, they mostly stay in the center of the ground truth and do not extend to the full region of the tract. Also, for the cases with worst performance, it is noticeable that they are a bit shifted from the center potentially affecting the measured overlap.

Table 7 – Averaged results for five-fold cross-validation in experiments 1 and 2. The best scores are emboldened and the worst are underlined.

Subjects	**Dice Coefficient (%)**	
	Experiment 1	**Experiment 2**
1	20.508 ± 0.032	22.07 ± 0.072
2	16.908 ± 0.007	21.133 ± 0.041
3	15.144 ± 0.018	19.925 ± 0.054
4	19.869 ± 0.025	23.894 ± 0.031
5	25.737 ± 0.060	$\mathbf{27.849 \pm 0.047}$
6	$\mathbf{28.598 \pm 0.076}$	24.206 ± 0.063
7	15.470 ± 0.047	$\underline{5.544 \pm 0.046}$
8	18.780 ± 0.071	18.856 ± 0.059
9	$\underline{12.807 \pm 0.04}$	21.044 ± 0.069
10	16.700 ± 0.060	23.094 ± 0.042

3.4 Discussion

3.4.1 Experiment 1

Experiment 1 consisted of a five-fold cross-validation experiment using data from the same subject for the training and testing stages. In this experiment, we aimed to investigate how well we predict tracts from the same dataset. Although the Dice coefficient (Table 7) is low, it is seem in the 3D reconstructions depicted in Figure 22 that subject 6 followed the fibers bundle reasonably. However, this pattern did not persist over all folds (Figure 22(a) and Figures 22(c)-(e)). Therefore, we can assume that the generalization of our method could be improved.

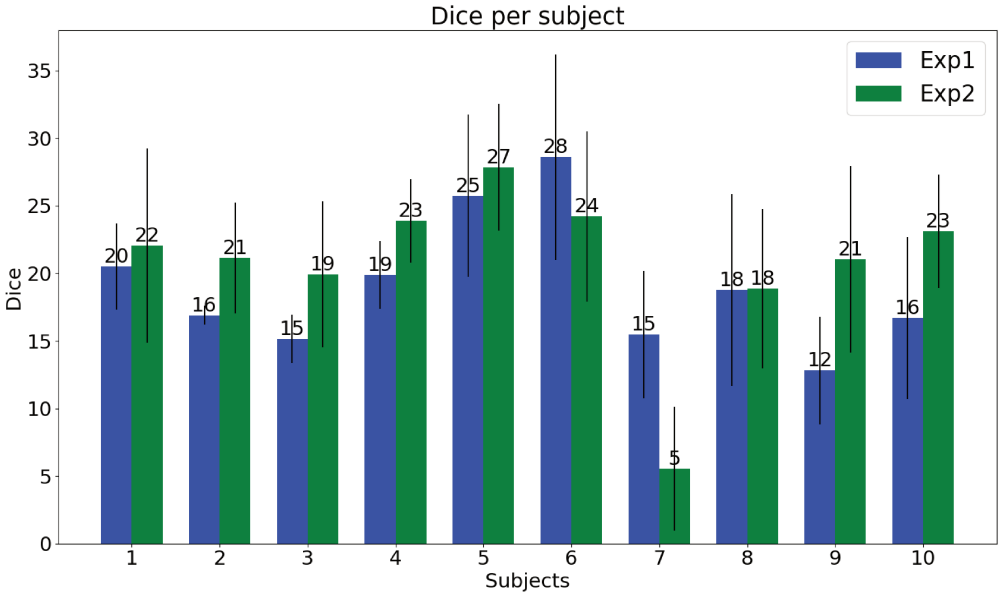


Figure 21 – Bar plot showing the overlap differences computed by the Dice in Experiment 1 and 2. The vertical lines in black are a measure of the standard deviation per subject along the experiments.

3.4.2 Experiment 2

In the second experiment, we aimed to investigate the inference our method by inserting variability into the training and testing sets with the use of fibers bundle from different subjects. The assumption is that the inter-subject variability of the UF tract could be beneficial in learning more generalized features for the fiber tracts. Nonetheless, the Dice coefficient performance did not significantly improve ($p-value = 0.35$). For most of the subjects, there was an enhancement in the overlap compared to Experiment 1. The predicted fibers bundles followed the ground truth as in Experiment 1, but were more centralized than the previous experiment (Figure 23). Moreover, the lack of generalization continued among the subjects and folds which needs further investigation.

3.4.3 General Discussion

In general, the tracts seem to be similar in length to the ground truth, and converge on the mean tract shape (Figures 22 and 23). Our approach shows, however, there still work to be done to ensure the fiber tract variation is captured when automatically extracting the fiber tracts.

The high complexity of the 3D ResNet might have affected the convergence during the training stage. In our CNN, we have 33M trainable parameters and even the 5000 input fibers from Experiment 1 may not be sufficient to train the model. Hence, the

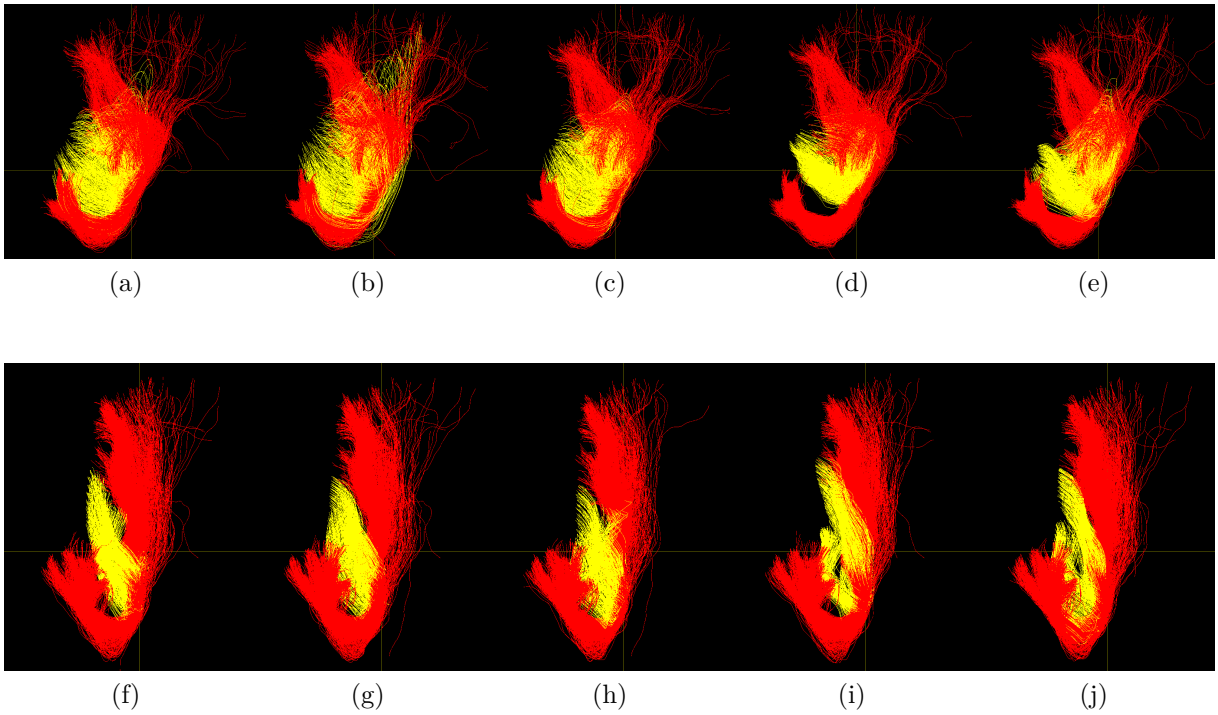


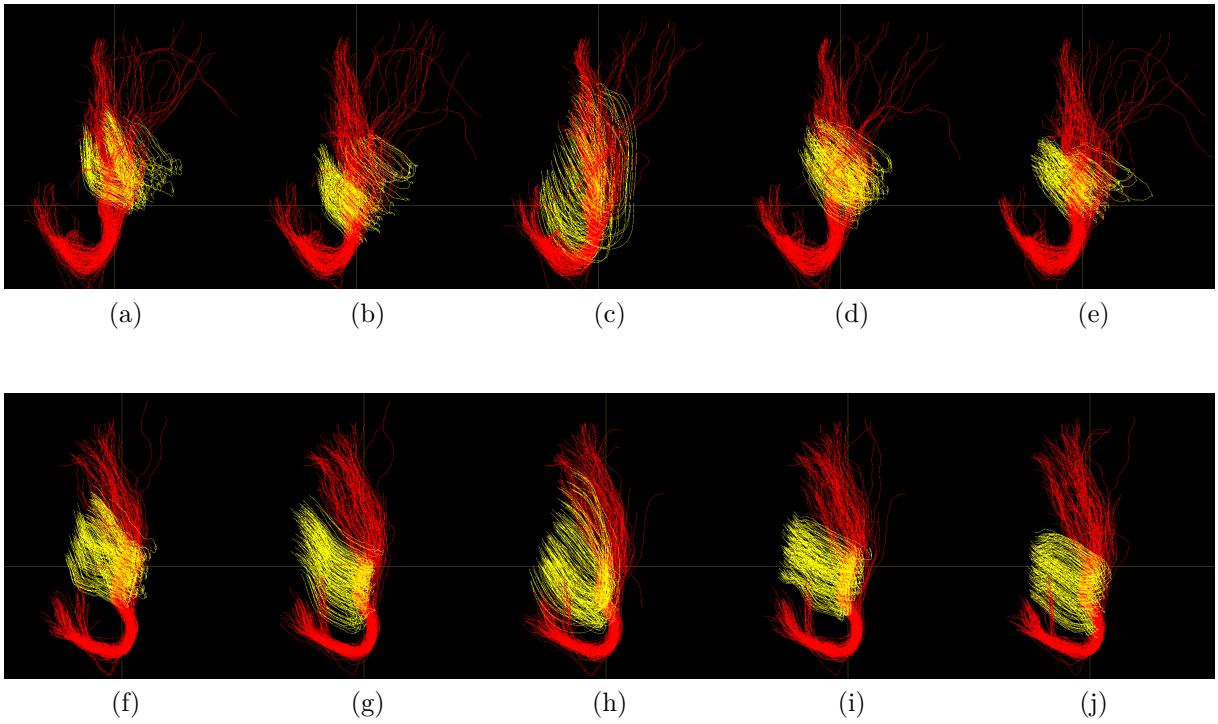
Figure 22 – Representative 3D reconstruction of the left UF tract in Experiment 1, which the ground truth annotation is in red and the prediction from our method is yellow. (a)-(e) 3D reconstructions of the subject **6** (best performance among all subjects) and (f)-(j) subject **9** (worst performance among all subjects)

network might not converge properly to a minimum. In Experiment 2, due to our memory resources, we could not store all the 5000 fibers from each patient, and the problem of obtaining enough data to fit the model persisted. Therefore, there is a necessity to leverage more data during training.

Even though the Huber loss is a robust cost function, it only measures a point-wise distance and does not take into account the geometry of the fibers. That may explain why the fibers were near the mean shape and not fitting the full length of the fiber tract. A possible solution to overcome this case is to find a cost function that computes the curve distance (MAHENDRAN *et al.*, 2017) or incorporating statistical shape priors (MILLETARI *et al.*, 2017).

3.5 Conclusions

In this chapter, we have proposed a deep-learning-based approach to perform tractography for surgical planning in epilepsy treatment. We performed analysis in a single patient and also among 10 patients in a cross validation approach. Although the results were not optimal, the tracts tended to be of a similar length and converged to the mean fiber tract locations. As fas as we know, our method is the first approach that investigates



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 23 – Representative 3D reconstruction of the left UF tract in Experiment 2, which the ground truth annotation is in red and the prediction from our method is yellow. (a)-(e) 3D reconstructions of the subject **5** (best performance among all subjects) and (f)-(j) subject **7** (worst performance among all subjects)

CNNs for tractography. Therefore, our work is a start point for further analysis regarding the presented topic. We recognize the limitations of our method, and as future work, we intend to acquire more training data, investigate the use of curve distance metrics for the cost function, and incorporate statistical shape analysis as priors.

4 Conclusions

This dissertation advanced the study of deep-learning-based approaches for brain MR analysis. This technique was investigated in a segmentation task for skull-stripping (SS) and in a regression analysis for tractography.

With respect to SS analysis, we first validated silver standard masks for CNN training. Then, a robust CNN for brain MR imaging SS was proposed. As far as we know, our approaches were the first ones fully trained using silver standards masks, and thus, eliminating the use of expert manual annotation. CONSNet and its auto-context version achieved performance comparable to SS state-of-the-art automatic methods, faster than the most robust non-deep-learning-based methods, and had an optimal generalization across datasets. Moreover, silver standard masks are a low-cost solution for generating annotated data for small datasets, and also, they reduce inter-/intra-rater variability and overcome the bias of data-driven approaches due to the consensus agreement among different automatic methods. These results aim to leverage the use of silver standard brain masks for large-scale studies in medical image processing.

Regarding this first task, there are challenges that can still be investigated. In our analyses, we implemented 2D CNN architectures in a tri-planar approach. However, a 3D analyses can improve the output due to the 3D neighborhood features. Therefore, we plan to extend CONSNet to a 3D architecture. Also, we recognize the paucity of expertly annotated data in other medical image processing tasks. Manual annotation is not only time-consuming for SS but is known to vary, even among highly-trained experts (WARFIELD *et al.*, 2004; AKKUS *et al.*, 2017), and be impacted by both inter- and intra-rater variability (ASMAN; LANDMAN, 2011). Hence, silver standard masks can impact other applications.

Concerning the deep-learning-based tractography for surgical planning in epilepsy treatment, our contribution to the field was a baseline approach for CNN-based tractography. Previous works in deep learning consisted of methods with no convolution layers. To the best of our knowledge, we were the first to propose a model with them. Our analyses were conducted using a cross-validation for a single patient, and also, a cross-validation using data from 10 subjects at the training stage. Additionally, we performed a regression based on a three-stage methodology with an auto-encoder, training a CNN from scratch, and a fine-tuning phase. Although the overlap metrics were low, the results showed tracts tended to be of a similar length and converged to the mean fiber tract locations.

Outputting a fiber from a tract in an end-to-end CNN solution is a complex task. Training 3D CNN models is costly and requires tones of data. For instance, in one of the experiments, we have 33M trainable parameters and even the 5000 input fibers used were

not sufficient to train the model properly. Therefore, there is a necessity to leverage more data during training. Another challenge is to find an appropriate loss function to take into account the geometry of the fiber. A possible solution to overcome this case is to find a cost function that computes curve distance (MAHENDRAN *et al.*, 2017) or incorporating statistical shape priors (MILLETARI *et al.*, 2017). We recognize that for this task our results were not optimal. However, a further investigation acquiring more training data, performing CNN model selection, and using a curve distance loss function can improve this baseline analysis.

4.1 Publications

As a result of this dissertation, one journal paper, two international conference full papers, and three abstracts were published. Remark that the abstracts were published based on preliminary findings for the skull-stripping case during the M.Sc. project. The conference paper *"Transfer Learning Using Convolutional Neural Networks for Face Anti-spoofing"* was published as a result of M.Sc. course project which I had a first contact with deep-learning- based techniques that helped me in further analysis for brain MRI skull-stripping. Additionally, the conference paper *"A machine learning approach to predict instrument bending in stereotactic neurosurgery"* was a result of a collaborative work during my M.Sc. internship at University College London, which I was responsible for the machine learning analysis. A full list of publications is presented as follows:

4.1.1 Journal Papers

- Souza, R., **Lucena, O.**, Garrafa, J., Gobbi, D., Saluzzi, M., Appenzeller, S., Rittner, L., Frayne, R., Lotufo, R., *An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement*, Neuroimage, 2017

4.1.2 Conference Papers

- **Lucena, O.**, Souza, R., Rittner, L., Frayne, R., Lotufo, R. (2018, April). *Silver standard masks for data augmentation applied to deep-learning-based skull-stripping.* 2018 IEEE 15th International Symposium on Biomedical Imaging (IEEE ISBI 2018).

- Souza, R., **Lucena, O**., Bento, M., Garrafa, J., Appenzeller, S., Rittner, L., Lotufo, R. and Frayne, R., 2018, April. *Reliability of using single specialist annotation for designing and evaluating automatic segmentation methods: A skull stripping case study.* 2018 IEEE 15th International Symposium on Biomedical Imaging (IEEE ISBI 2018)

- **Lucena, O.**, Junior, A., Moia, V., Souza, R., Valle, E., Lotufo, R., *"Transfer Learning Using Convolutional Neural Networks for Face Anti-spoofing"*, 14th International Conference Image Analysis and Recognition (ICIAR 2017).

- Granados, A., Mancini, M., Vos, S., **Lucena, O.**, Vakharia, V., Rodionov, R., Miserocchi, A., McEvoy, A., Duncan, J., Sparks, R., Ourselin, S., *A machine learning approach to predict instrument bending in stereotactic neurosurgery.* 21st International Conference On Medical Image Computing and Computer Assisted Intervention (MICCAI 2018).

4.1.3 Abstracts

- **Lucena, O.**, Souza, R., Frayne, R., Rittner, L., Lotufo. *2D Single Plane Big data Convolutional Neural Network for skull stripping.* 27th ISMRM 2018.

- Souza, R., **Lucena, O.**, Rittner, L., Lotufo, R., Frayne, R. *Can brain MRI skull-stripping methods be further improved using manual segmentation as ground-truth for validation?* 27th ISMRM 2018.

- **Lucena, O.**, Souza, R., Lotufo, R. *"A 2D CNN Approach for Skull-Stripping in MR Imaging"*. 4th BRAINN Congress, 2017, Journal of Epilepsy and Clinical Neurophysiology, 2017.

Acknowledgments

I would like to thank the São Paulo Research Foundation (FAPESP) processes 2016/18332-8 and 2017/23747-5 for providing financial support.

Bibliography

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M. *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. Citado 2 vezes nas páginas 26 and 53.

AKKUS, Z.; GALIMZIANOVA, A.; HOOGI, A.; RUBIN, D. L.; ERICKSON, B. J. Deep learning for brain mri segmentation: State of the art and future directions. *Journal of Digital Imaging*, Springer, p. 1–11, 2017. Citado 2 vezes nas páginas 20 and 62.

ALEXANDER, A. L. Deterministic white matter tractography. *Diffusion MRI: Theory, methods, and applications*, Oxford University Press USA, p. 383–395, 2010. Citado na página 51.

ALJABAR, P.; HECKEMANN, R. A.; HAMMERS, A.; HAJNAL, J. V.; RUECKERT, D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage*, Elsevier, v. 46, n. 3, p. 726–738, 2009. Citado na página 20.

ANDERSSON, J. L.; SKARE, S.; ASHBURNER, J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*, Elsevier, v. 20, n. 2, p. 870–888, 2003. Citado na página 52.

ASMAN, A. J.; LANDMAN, B. A. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (collate). *IEEE Trans. Med. Imag.*, IEEE, v. 30, n. 10, p. 1779–1794, 2011. Citado 3 vezes nas páginas 20, 25, and 62.

ASSAF, Y.; BLUMENFELD-KATZIR, T.; YOVEL, Y.; BASSER, P. J. Axcaliber: a method for measuring axon diameter distribution from diffusion mri. *Magnetic resonance in medicine*, Wiley Online Library, v. 59, n. 6, p. 1347–1354, 2008. Citado na página 51.

AVANTS, B. B.; EPSTEIN, C. L.; GROSSMAN, M.; GEE, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, Elsevier, v. 12, n. 1, p. 26–41, 2008. Citado na página 56.

AVANTS, B. B.; TUSTISON, N. J.; SONG, G.; COOK, P. A.; KLEIN, A.; GEE, J. C. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, Elsevier, v. 54, n. 3, p. 2033–2044, 2011. Citado 4 vezes nas páginas 19, 21, 24, and 31.

BAR, Y.; DIAMANT, I.; WOLF, L.; GREENSPAN, H. Deep learning with non-medical training used for chest pathology identification. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Medical Imaging 2015: Computer-Aided Diagnosis*. [S.l.], 2015. v. 9414, p. 94140V. Citado na página 16.

BASSER, P. J.; MATTIELLO, J.; LEBIHAN, D. Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, Elsevier, v. 66, n. 1, p. 259–267, 1994. Citado na página 51.

BEARE, R.; CHEN, J.; ADAMSON, C. L.; SILK, T.; THOMPSON, D. K.; YANG, J. Y.; ANDERSON, V. A.; SEAL, M. L.; WOOD, A. G. Brain extraction using the watershed transform from markers. *Frontiers in neuroinformatics*, Frontiers Media SA, v. 7, 2013. Citado 3 vezes nas páginas 19, 20, and 24.

BERMAN, J. Diffusion mr tractography as a tool for surgical planning. *Magnetic resonance imaging clinics of North America*, Elsevier, v. 17, n. 2, p. 205–214, 2009. Citado 3 vezes nas páginas 17, 50, and 51.

BLANTON, R. E.; LEVITT, J. G.; PETERSON, J. R.; FADALE, D.; SPORTY, M. L.; LEE, M.; TO, D.; MORMINO, E. C.; THOMPSON, P. M.; MCCRACKEN, J. T. *et al.* Gender differences in the left inferior frontal gyrus in normal children. *Neuroimage*, Elsevier, v. 22, n. 2, p. 626–636, 2004. Citado na página 19.

BOCCARDI, M.; GANZOLA, R.; BOCCHETTA, M.; PIEVANI, M.; REDOLFI, A.; BARTZOKIS, G.; CAMICIOLI, R.; CSERNANSKY, J. G.; LEON, M. J. D.; DETOLEDO-MORRELL, L. *et al.* Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint eadc-adni harmonized protocol. *Journal of Alzheimer's Disease*, IOS Press, v. 26, n. s3, p. 61–75, 2011. Citado na página 20.

BOER, R. de; VROOMAN, H. A.; IKRAM, M. A.; VERNOOIJ, M. W.; BRETELER, M. M.; LUGT, A. van der; NIESSEN, W. J. Accuracy and reproducibility study of automatic mri brain tissue segmentation methods. *Neuroimage*, Elsevier, v. 51, n. 3, p. 1047–1056, 2010. Citado na página 19.

BREBISSON, A. de; MONTANA, G. Deep neural networks for anatomical brain segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2015. p. 20–28. Citado na página 16.

BROWN, R. W.; CHENG, Y.-C. N.; HAACKE, E. M.; THOMPSON, M. R.; VENKATESAN, R. *Magnetic resonance imaging: physical principles and sequence design*. [S.l.]: John Wiley & Sons, 2014. Citado na página 16.

CAN, A.; BELLO, M.; CLINE, H. E.; TAO, X.; MENDONCA, P.; GERDES, M. A unified segmentation method for detecting subcellular compartments in immunofluroescently labeled tissue images. In: *International Workshop on Microscopic Image Analysis with Applications in Biology. Sept.* [S.l.: s.n.], 2009. v. 3. Citado na página 25.

CHAKRAVARTY, M. M.; STEADMAN, P.; EEDE, M. C.; CALCOTT, R. D.; GU, V.; SHAW, P.; RAZNAHAN, A.; COLLINS, D. L.; LERCH, J. P. Performing label-fusion-based segmentation using multiple automatically generated templates. *Human brain mapping*, Wiley Online Library, v. 34, n. 10, p. 2635–2654, 2013. Citado na página 27.

CHEN, H.; DOU, Q.; YU, L.; QIN, J.; HENG, P.-A. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, Elsevier, 2017. Citado 2 vezes nas páginas 21 and 28.

CHENG, X.; ZHANG, L.; ZHENG, Y. Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis, v. 6, n. 3, p. 248–252, 2018. Citado na página 16.

DOZAT, T. Incorporating nesterov momentum into adam. 2016. Citado na página 53.

ESKILDSEN, S. F.; COUPÉ, P.; FONOV, V.; MANJÓN, J. V.; LEUNG, K. K.; GUIZARD, N.; WASSEF, S. N.; ØSTERGAARD, L. R.; COLLINS, D. L.; INITIATIVE, A. D. N. *et al.* Beast: brain extraction based on nonlocal segmentation technique. *NeuroImage*, Elsevier, v. 59, n. 3, p. 2362–2373, 2012. Citado 4 vezes nas páginas 19, 21, 24, and 47.

ESSAYED, W. I.; ZHANG, F.; UNADKAT, P.; COSGROVE, G. R.; GOLBY, A. J.; O'DONNELL, L. J. White matter tractography for neurosurgical planning: A topography-based review of the current state of the art. *NeuroImage: Clinical*, Elsevier, v. 15, p. 659–672, 2017. Citado na página 54.

FENNEMA-NOTESTINE, C.; OZYURT, I. B.; CLARK, C. P.; MORRIS, S.; BISCHOFF-GRETHE, A.; BONDI, M. W.; JERNIGAN, T. L.; FISCHL, B.; SEGONNE, F.; SHATTUCK, D. W. *et al.* Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location. *Human brain mapping*, Wiley Online Library, v. 27, n. 2, p. 99–113, 2006. Citado na página 25.

FISHER, R. S.; BOAS, W. v. E.; BLUME, W.; ELGER, C.; GENTON, P.; LEE, P.; ENGEL, J. Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, Wiley Online Library, v. 46, n. 4, p. 470–472, 2005. Citado na página 50.

GOONERATNE, I. K.; MANNAN, S.; TISI, J. de; GONZALEZ, J. C.; MCEVOY, A. W.; MISEROCCHI, A.; DIEHL, B.; WEHNER, T.; BELL, G. S.; SANDER, J. W. *et al.* Somatic complications of epilepsy surgery over 25 years at a single center. *Epilepsy research*, Elsevier, v. 132, p. 70–77, 2017. Citado na página 50.

GREENSPAN, H.; GINNEKEN, B. van; SUMMERS, R. M. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Trans. Med. Imag.*, IEEE, v. 35, n. 5, p. 1153–1159, 2016. Citado 2 vezes nas páginas 16 and 19.

HAMMERNIK, K.; KLATZER, T.; KOBLER, E.; RECHT, M. P.; SODICKSON, D. K.; POCK, T.; KNOLL, F. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, Wiley Online Library, v. 79, n. 6, p. 3055–3071, 2018. Citado na página 16.

HAYNES, W. Wilcoxon rank sum test. In: *Encyclopedia of Systems Biology*. [S.l.]: Springer, 2013. p. 2354–2355. Citado na página 31.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado 2 vezes nas páginas 21 and 53.

HUANG, X.; SHAN, J.; VAIDYA, V. Lung nodule detection in ct using 3d convolutional neural networks. In: IEEE. *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. [S.l.], 2017. p. 379–383. Citado na página 16.

HUFF, T. J.; LUDWIG, P. E.; ZUNIGA, J. M. The potential for machine learning algorithms to improve and reduce the cost of 3-dimensional printing for surgical planning. *Expert review of medical devices*, Taylor & Francis, v. 15, n. 5, p. 349–356, 2018. Citado na página 17.

HUTCHINSON, M.; RAFF, U. Structural changes of the substantia nigra in parkinson's disease as revealed by mr imaging. *American journal of neuroradiology*, Am Soc Neuroradiology, v. 21, n. 4, p. 697–701, 2000. Citado na página 19.

IBRAGIMOV, B.; TOESCA, D.; CHANG, D.; KOONG, A.; XING, L. Combining deep learning with anatomical analysis for segmentation of the portal vein for liver sbrt planning. *Physics in Medicine & Biology*, IOP Publishing, v. 62, n. 23, p. 8943, 2017. Citado na página 17.

IGLESIAS, J. E.; LIU, C.-Y.; THOMPSON, P. M.; TU, Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. Med. Imag.*, IEEE, v. 30, n. 9, p. 1617–1634, 2011. Citado 3 vezes nas páginas 19, 21, and 24.

JEURISSEN, B.; LEEMANS, A.; JONES, D. K.; TOURNIER, J.-D.; SIJBERS, J. Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution. *Human brain mapping*, Wiley Online Library, v. 32, n. 3, p. 461–479, 2011. Citado na página 51.

JOHNSON, H. J.; MCCORMICK, M. M.; IBANEZ, L. *The ITK Software Guide Book 1: Introduction and Development Guidelines - Volume 1*. USA: Kitware, Inc., 2015. ISBN 1930934270, 9781930934276. Citado na página 26.

KALAVATHI, P.; PRASATH, V. S. Methods on skull stripping of mri head scan images—a review. *Journal of digital imaging*, Springer, v. 29, n. 3, p. 365–379, 2016. Citado 3 vezes nas páginas 16, 19, and 20.

KAMNITSAS, K.; LEDIG, C.; NEWCOMBE, V. F.; SIMPSON, J. P.; KANE, A. D.; MENON, D. K.; RUECKERT, D.; GLOCKER, B. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, Elsevier, v. 36, p. 61–78, 2017. Citado na página 21.

KATIRCIOGLU, I.; TEKIN, B.; SALZMANN, M.; LEPETIT, V.; FUA, P. Learning latent representations of 3d human pose with deep neural networks. *International Journal of Computer Vision*, Springer, p. 1–16, 2018. Citado na página 54.

KER, J.; WANG, L.; RAO, J.; LIM, T. Deep learning applications in medical image analysis. *IEEE Access*, IEEE, v. 6, p. 9375–9389, 2018. Citado na página 16.

KINGMA, D.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Citado na página 54.

KLAMBAUER, G.; UNTERTHINER, T.; MAYR, A.; HOCHREITER, S. Self-normalizing neural networks. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2017. p. 972–981. Citado na página 53.

KLEESIEK, J.; URBAN, G.; HUBERT, A.; SCHWARZ, D.; MAIER-HEIN, K.; BENDSZUS, M.; BILLER, A. Deep mri brain extraction: a 3d convolutional neural

network for skull stripping. *NeuroImage*, Elsevier, v. 129, p. 460–469, 2016. Citado 9 vezes nas páginas , 19, 21, 24, 25, 31, 32, 45, and 46.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105. Citado na página 19.

KWAN, P.; BRODIE, M. J. Drug treatment of epilepsy: when does it fail and how to optimize its use? *CNS spectrums*, Cambridge University Press, v. 9, n. 2, p. 110–119, 2004. Citado na página 50.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Research, v. 521, n. 7553, p. 436–444, 2015. Citado 3 vezes nas páginas 16, 19, and 52.

LEMKADDEM, A.; SKIÖLDEBRAND, D.; PALÚ, A. D.; THIRAN, J.-P.; DADUCCI, A. Global tractography with embedded anatomical priors for quantitative connectivity analysis. *Frontiers in neurology*, Frontiers Media SA, v. 5, 2014. Citado 2 vezes nas páginas 50 and 51.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. A. van der; GINNEKEN, B. van; SÁNCHEZ, C. I. A survey on deep learning in medical image analysis. *arXiv preprint arXiv:1702.05747*, 2017. Citado 2 vezes nas páginas 16 and 19.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2015. p. 3431–3440. Citado na página 21.

LUCENA, O.; SOUZA, R.; RITTNER, L.; FRAYNE, R.; LOTUFO, R. Silver standard masks for data augmentation applied to deep-learning-based skull-stripping. In: IEEE. *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. [S.l.], 2018. p. 1114–1117. Citado 2 vezes nas páginas 16 and 22.

LÜDERS, H. O.; NAJM, I.; NAIR, D.; WIDDESS-WALSH, P.; BINGMAN, W. The epileptogenic zone: general principles. *Epileptic disorders*, Citeseer, v. 8, n. 2, p. 1–9, 2006. Citado 2 vezes nas páginas 17 and 50.

LUTKENHOFF, E. S.; ROSENBERG, M.; CHIANG, J.; ZHANG, K.; PICKARD, J. D.; OWEN, A. M.; MONTI, M. M. Optimized brain extraction for pathological brains (optibet). *PLoS One*, Public Library of Science, v. 9, n. 12, p. e115551, 2014. Citado 2 vezes nas páginas 21 and 24.

MAHENDRAN, S.; ALI, H.; VIDAL, R. 3d pose regression using convolutional neural networks. In: *IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. v. 1, n. 2, p. 4. Citado 2 vezes nas páginas 60 and 63.

MAIER-HEIN, K. H.; NEHER, P. F.; HOUDE, J.-C.; CÔTÉ, M.-A.; GARYFALLIDIS, E.; ZHONG, J.; CHAMBERLAND, M.; YEH, F.-C.; LIN, Y.-C.; JI, Q. *et al.* The challenge of mapping the human connectome based on diffusion tractography. *Nature communications*, Nature Publishing Group, v. 8, n. 1, p. 1349, 2017. Citado na página 54.

MALCOLM, J. G.; SHENTON, M. E.; RATHI, Y. Filtered multitensor tractography. *IEEE transactions on medical imaging*, IEEE, v. 29, n. 9, p. 1664–1675, 2010. Citado na página 51.

MARCUS, D. S.; WANG, T. H.; PARKER, J.; CSERNANSKY, J. G.; MORRIS, J. C.; BUCKNER, R. L. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, MIT Press, v. 19, n. 9, p. 1498–1507, 2007. Citado 2 vezes nas páginas 22 and 24.

MENEGOLA, A.; TAVARES, J.; FORNACIALI, M.; LI, L. T.; AVILA, S.; VALLE, E. Recod titans at isic challenge 2017. *arXiv preprint arXiv:1703.04819*, 2017. Citado na página 26.

MILLETARI, F.; ROTHBERG, A.; JIA, J.; SOFKA, M. Integrating statistical prior knowledge into convolutional neural networks. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2017. p. 161–168. Citado 2 vezes nas páginas 60 and 63.

MORI, S.; CRAIN, B. J.; CHACKO, V.; ZIJL, P. V. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of neurology*, Wiley Online Library, v. 45, n. 2, p. 265–269, 1999. Citado na página 51.

NEHER, P. F.; COTE, M.-A.; HOUDE, J.-C.; DESCOTEAUX, M.; MAIER-HEIN, K. H. Fiber tractography using machine learning. *NeuroImage*, v. 158, p. 417 – 429, 2017. Citado 3 vezes nas páginas 17, 51, and 52.

NEHER, P. F.; DESCOTEAUX, M.; HOUDE, J.-C.; STIELTJES, B.; MAIER-HEIN, K. H. Strengths and weaknesses of state of the art fiber tractography pipelines–a comprehensive in-vivo and phantom evaluation study using tractometer. *Medical image analysis*, Elsevier, v. 26, n. 1, p. 287–305, 2015. Citado na página 51.

NOWELL, M.; VOS, S. B.; SIDHU, M.; WILCOXEN, K.; SARGSYAN, N.; OURSELIN, S.; DUNCAN, J. S. Meyer's loop asymmetry and language lateralisation in epilepsy. *J Neurol Neurosurg Psychiatry*, BMJ Publishing Group Ltd, p. jnnp–2015, 2015. Citado na página 51.

PARKER, G. Probabilistic fiber tracking. *Diffusion MRI: Theory, methods, and applications*, Oxford University Press, p. 396–408, 2010. Citado na página 51.

PETRELLA, J. R.; COLEMAN, R. E.; DORAISWAMY, P. M. Neuroimaging and early diagnosis of alzheimer disease: a look to the future. *Radiology*, Radiological Society of North America, v. 226, n. 2, p. 315–336, 2003. Citado na página 19.

PLATT, M.; SPERLING, M. R. A comparison of surgical and medical costs for refractory epilepsy. *Epilepsia*, Wiley Online Library, v. 43, n. s4, p. 25–31, 2002. Citado na página 50.

POULIN, P.; COTE, M.-A.; HOUDE, J.-C.; PETIT, L.; NEHER, P. F.; MAIER-HEIN, K. H.; LAROCHELLE, H.; DESCOTEAUX, M. Learn to track: Deep learning for tractography. *bioRxiv*, Cold Spring Harbor Labs Journals, p. 146688, 2017. Citado 2 vezes nas páginas 17 and 51.

PRASOON, A.; PETERSEN, K.; IGEL, C.; LAUZE, F.; DAM, E.; NIELSEN, M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: SPRINGER. *International conference on medical image computing and computer-assisted intervention*. [S.l.], 2013. p. 246–253. Citado na página 26.

RAVÌ, D.; WONG, C.; DELIGIANNI, F.; BERTHELOT, M.; ANDREU-PEREZ, J.; LO, B.; YANG, G.-Z. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, IEEE, v. 21, n. 1, p. 4–21, 2017. Citado na página 16.

REHM, K.; SCHAPER, K.; ANDERSON, J.; WOODS, R.; STOLTZNER, S.; ROTTENBERG, D. Putting our heads together: a consensus approach to brain/non-brain segmentation in t1-weighted mr volumes. *NeuroImage*, Elsevier, v. 22, n. 3, p. 1262–1270, 2004. Citado na página 25.

REISERT, M.; MADER, I.; ANASTASOPOULOS, C.; WEIGEL, M.; SCHNELL, S.; KISELEV, V. Global fiber reconstruction becomes practical. *Neuroimage*, Elsevier, v. 54, n. 2, p. 955–962, 2011. Citado na página 51.

REX, D. E.; SHATTUCK, D. W.; WOODS, R. P.; NARR, K. L.; LUDERS, E.; REHM, K.; STOLZNER, S. E.; ROTTENBERG, D. A.; TOGA, A. W. A meta-algorithm for brain extraction in mri. *NeuroImage*, Elsevier, v. 23, n. 2, p. 625–637, 2004. Citado 3 vezes nas páginas 20, 21, and 25.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2015. p. 234–241. Citado 2 vezes nas páginas 21 and 26.

RUSINEK, H.; LEON, M. J. de; GEORGE, A. E.; STYLOPOULOS, L.; CHANDRA, R.; SMITH, G.; RAND, T.; MOURINO, M.; KOWALSKI, H. Alzheimer disease: measuring loss of cerebral gray matter with mr imaging. *Radiology*, v. 178, n. 1, p. 109–114, 1991. Citado na página 19.

SALEHI, S. S. M.; ERDOGMUS, D.; GHOLIPOUR, A. Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE Trans. Med. Imag.*, IEEE, 2017. Citado 10 vezes nas páginas , 19, 21, 24, 25, 31, 32, 45, 46, and 48.

SALEMBIER, P.; OLIVERAS, A.; GARRIDO, L. Antiextensive connected operators for image and sequence processing. *IEEE Trans. Image Process.*, IEEE, v. 7, n. 4, p. 555–570, 1998. Citado na página 30.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural networks*, Elsevier, v. 61, p. 85–117, 2015. Citado na página 16.

SÉGONNE, F.; DALE, A. M.; BUSA, E.; GLESSNER, M.; SALAT, D.; HAHN, H. K.; FISCHL, B. A hybrid approach to the skull stripping problem in mri. *Neuroimage*, Elsevier, v. 22, n. 3, p. 1060–1075, 2004. Citado 3 vezes nas páginas 19, 21, and 24.

SHATTUCK, D. W.; MIRZA, M.; ADISETIYO, V.; HOJATKASHANI, C.; SALAMON, G.; NARR, K. L.; POLDRACK, R. A.; BILDER, R. M.; TOGA, A. W. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage*, Elsevier, v. 39, n. 3, p. 1064–1080, 2008. Citado 2 vezes nas páginas 22 and 24.

SHATTUCK, D. W.; SANDOR-LEAHY, S. R.; SCHAPER, K. A.; ROTTENBERG, D. A.; LEAHY, R. M. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, Elsevier, v. 13, n. 5, p. 856–876, 2001. Citado 3 vezes nas páginas 19, 20, and 24.

SMITH, S. M. Fast robust automated brain extraction. *Human brain mapping*, Wiley Online Library, v. 17, n. 3, p. 143–155, 2002. Citado 3 vezes nas páginas 19, 20, and 24.

SOUZA, R.; LUCENA, O.; BENTO, M.; GARRAFA, J.; APPENZELLER, S.; RITTNER, L.; LOTUFO, R.; FRAYNE, R. Reliability of using single specialist annotation for designing and evaluating automatic segmentation methods: A skull stripping case study. In: IEEE. *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. [S.l.], 2018. p. 1344–1347. Citado na página 16.

SOUZA, R.; LUCENA, O.; GARRAFA, J.; GOBBI, D.; SALUZZI, M.; APPENZELLER, S.; RITTNER, L.; FRAYNE, R.; LOTUFO, R. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, Elsevier, 2017. Citado 2 vezes nas páginas 22 and 23.

TANSKANEN, P.; VEIJOLA, J. M.; PIIPPO, U. K.; HAAPEA, M.; MIETTUNEN, J. A.; PYHTINEN, J.; BULLMORE, E. T.; JONES, P. B.; ISOHANNI, M. K. Hippocampus and amygdala volumes in schizophrenia and other psychoses in the northern finland 1966 birth cohort. *Schizophrenia research*, Elsevier, v. 75, n. 2, p. 283–294, 2005. Citado na página 19.

TIELEMAN, T.; HINTON, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, v. 4, n. 2, p. 26–31, 2012. Citado na página 26.

TOURNIER, J.; CALAMANTE, F.; CONNELLY, A. *et al.* Mrtrix: diffusion tractography in crossing fiber regions. *International Journal of Imaging Systems and Technology*, Wiley Online Library, v. 22, n. 1, p. 53–66, 2012. Citado 3 vezes nas páginas 51, 53, and 56.

TOURNIER, J. D.; CALAMANTE, F.; CONNELLY, A. Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions. In: *Proc. 18th Annual Meeting of the Intl. Soc. Mag. Reson. Med.(ISMRM)*. [S.l.: s.n.], 2010. p. 1670. Citado na página 53.

TOURNIER, J.-D.; CALAMANTE, F.; GADIAN, D. G.; CONNELLY, A. Direct estimation of the fiber orientation density function from diffusion-weighted mri data using spherical deconvolution. *NeuroImage*, Elsevier, v. 23, n. 3, p. 1176–1185, 2004. Citado na página 51.

TU, Z.; BAI, X. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell."*, IEEE, v. 32, n. 10, p. 1744–1757, 2010. Citado 2 vezes nas páginas 21 and 28.

WAKANA, S.; CAPRIHAN, A.; PANZENBOECK, M. M.; FALLON, J. H.; PERRY, M.; GOLLUB, R. L.; HUA, K.; ZHANG, J.; JIANG, H.; DUBEY, P. *et al.* Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage*, Elsevier, v. 36, n. 3, p. 630–644, 2007. Citado na página 53.

WANG, G.; ZULUAGA, M. A.; LI, W.; PRATT, R.; PATEL, P. A.; AERTSEN, M.; DOEL, T.; DIVID, A. L.; DEPREST, J.; OURSELIN, S. *et al.* Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2018. Citado na página 17.

WARFIELD, S. K.; ZOU, K. H.; WELLS, W. M. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imag.*, IEEE, v. 23, n. 7, p. 903–921, 2004. Citado 5 vezes nas páginas 17, 20, 22, 25, and 62.

WASSERTHAL, J.; NEHER, P.; MAIER-HEIN, K. H. Tractseg-fast and accurate white matter tract segmentation. *arXiv preprint arXiv:1805.07103*, 2018. Citado 2 vezes nas páginas 54 and 55.

WIEBE, S.; BLUME, W. T.; GIRVIN, J. P.; ELIASZIW, M. A randomized, controlled trial of surgery for temporal-lobe epilepsy. *New England Journal of Medicine*, Mass Medical Soc, v. 345, n. 5, p. 311–318, 2001. Citado na página 50.

WINSTON, G. P.; DAGA, P.; STRETTON, J.; MODAT, M.; SYMMS, M. R.; MCEVOY, A. W.; OURSELIN, S.; DUNCAN, J. S. Optic radiation tractography and vision in anterior temporal lobe resection. *Annals of neurology*, Wiley Online Library, v. 71, n. 3, p. 334–341, 2012. Citado na página 50.

WINSTON, G. P.; MANCINI, L.; STRETTON, J.; ASHMORE, J.; SYMMS, M. R.; DUNCAN, J. S.; YOUSRY, T. A. Diffusion tensor imaging tractography of the optic radiation for epilepsy surgical planning: a comparison of two methods. *Epilepsy research*, Elsevier, v. 97, n. 1, p. 124–132, 2011. Citado na página 50.

WU, G.; WANG, Q.; ZHANG, D.; NIE, F.; HUANG, H.; SHEN, D. A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Medical image analysis*, Elsevier, v. 18, n. 6, p. 881–890, 2014. Citado na página 20.

YAMADA, K.; SAKAI, K.; AKAZAWA, K.; YUEN, S.; NISHIMURA, T. Mr tractography: a review of its clinical applications. *Magnetic resonance in medical sciences*, Japanese Society for Magnetic Resonance in Medicine, v. 8, n. 4, p. 165–174, 2009. Citado 3 vezes nas páginas 17, 50, and 51.

YENDIKI, A.; PANNECK, P.; SRINIVASAN, P.; STEVENS, A.; ZÖLLEI, L.; AUGUSTINACK, J.; WANG, R.; SALAT, D.; EHRLICH, S.; BEHRENS, T. *et al.* Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. *Frontiers in neuroinformatics*, Frontiers Media SA, v. 5, 2011. Citado na página 51.

ZHAN, L.; ZHOU, J.; WANG, Y.; JIN, Y.; JAHANSHAD, N.; PRASAD, G.; NIR, T. M.; LEONARDO, C. D.; YE, J.; THOMPSON, P. M. *et al.* Comparison of nine tractography algorithms for detecting abnormal structural brain networks in alzheimer's disease. *Frontiers in aging neuroscience*, Frontiers, v. 7, p. 48, 2015. Citado na página 51.

ZINKEVICH, M.; WEIMER, M.; LI, L.; SMOLA, A. J. Parallelized stochastic gradient descent. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2010. p. 2595–2603. Citado na página 54.

ZIVADINOV, R.; BAGNATO, F.; NASUELLI, D.; BASTIANELLO, S.; BRATINA, A.; LOCATELLI, L.; WATTS, K.; FINAMORE, L.; GROP, A.; DWYER, M. *et al.* Short-term brain atrophy changes in relapsing–remitting multiple sclerosis. *Journal of the neurological sciences*, Elsevier, v. 223, n. 2, p. 185–193, 2004. Citado na página 19.