UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

William Javier Garcia Herrera

Automatic quality control of corpus callosum segmentations in large population studies: comparing deep and classical learning

Controle de qualidade automático de segmentações do corpo caloso em grandes estudos populacionais: comparando abordagens de aprendizado profundo e clássico

Campinas

2020

William Javier Garcia Herrera

Automatic quality control of corpus callosum segmentations in large population studies: comparing deep and classical learning

Controle de qualidade automático de segmentações do corpo caloso em grandes estudos populacionais: comparando abordagens de aprendizado profundo e clássico

Thesis 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 Doctor of Philosophy in Electrical Engineering, in the field of Computer Engineering.

Tese 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 Doutor em Engenharia Elétrica, na Área de Engenharia de Computação.

Supervisor: Prof. Dr. Leticia Rittner

Este trabalho corresponde à versão final da tese defendida pelo aluno William Javier Garcia Herrera, e orientada pela Prof. Dr. Leticia Rittner.

Campinas

2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

G165a

Garcia Herrera, William Javier, 1985-
Automatic quality control of corpus callosum segmentations in large population studies : comparing deep and classical learning / William Javier Garcia Herrera. – Campinas, SP : [s.n.], 2020.

Orientador: Leticia Rittner.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Corpo caloso. 2. Imagens de ressonância magnética. 3. Controle de qualidade. 4. Máquina de vetores suporte. 5. Redes neurais convolucionais. I. Rittner, Leticia, 1972-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Controle de qualidade automático de segmentações do corpo caloso em grandes estudos populacionais : comparando abordagens de aprendizado profundo e clássico
Palavras-chave em inglês:
Corpus callosum
Magnetic resonance imaging
Quality control
Support vector machine
Convolutional neural network
Área de concentração: Engenharia de Computação
Titulação: Doutor em Engenharia Elétrica
Banca examinadora:
Leticia Rittner [Orientador]
Fernando José Von Zuben
Sandra Eliza Fontes de Avila
Carlos Ernesto Garrido Salmon
João Paulo Papa
Data de defesa: 22-06-2020
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-0429-8030
- Currículo Lattes do autor: http://lattes.cnpq.br/2959262851860061

COMISSÃO JULGADORA - TESE DE DOUTORADO

Candidato: William Javier Garcia Herrera RA: 162642
Data de defensa: 22 de junho de 2020
Thesis title: "Automatic quality control of corpus callosum segmentations in large population studies: comparing deep and classical learning"
Titulo da Tese: "Controle de qualidade automático de segmentações do corpo caloso em grandes estudos populacionais: comparando abordagens de aprendizado profundo e clássico"

Profa. Dra. Leticia Rittner (Presidente)
Prof. Dr. Fernando José Von Zuben (FEEC/Unicamp)
Profa. Dra. Sandra Eliza Fontes de Avila (IC/Unicamp)
Prof. Dr. Carlos Ernesto Garrido Salmon (USP)
Prof. Dr. João Paulo Papa (UNESP)

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Acknowledgements

This work was supported by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001, and the São Paulo Research Foundation (FAPESP - process CEPID 2013/07559-3) and by the National Council of Technological and Scientific Development (processes 190557/2014-1 and 308311/2016-7).

"Así has tomado el barro de mis años
Amasado de rabia y desengaños
para darle la forma verdadera"
(Fernando Soto Aparicio)

Abstract

The corpus callosum (CC) is the largest white matter structure in the central nervous system, allowing communication between both brain cortical hemispheres. This structure is important since its shape and volume are associated with certain subject characteristics, some diseases, and clinical conditions. Usually, the CC is studied in magnetic resonance imaging (MRI), where it is segmented to extract precise information and perform posterior analyses. As the availability of MRI data grows up and the automated algorithms to perform CC segmentation proliferate, quality control (QC) verification is mandatory to assure reliability in segmentations and avoid errors that would otherwise propagate throughout the analysis.

In this work, we compare two methods to perform QC of CC segmentations on T_1-MRI with no need for ground-truth. The first method uses classical machine learning techniques by performing manual extraction of a multi-resolution shape descriptor along with a classifier based on a support vector machine ensemble. The second method involves a deep learning approach by using a convolutional neural network (CNN) to extract deep features from the segmentation mask and contextual information from the image. In the experiments, images from 907 subjects were acquired from two sequences: 802 T_1-MRI (including 7 subjects with tumor affecting the CC), and 105 diffusion MRI (processed into fractional anisotropy maps). The dataset comes from *Hospital das Clínicas* from Unicamp, except for 247 T_1-MRI that come from ABIDE for the study of autism.

Both approaches got a similar performance (area under the curve ROC of 98%) when trained and tested on T_1-MRI as well as an identical execution time (9 seconds to process 136 samples). The performance of the methods was evaluated on two other datasets: diffusion MRI and patients with tumor. In diffusion MRI, the classical approach presented the best performance (AUC was 20% higher). In a dataset of patients with tumor affecting the CC, the CNN prevailed with an accuracy of 80%. The CNN was more versatile to learn new shapes and image intensities.

Keywords: corpus callosum; magnetic resonance imaging; quality control; support vector machine; convolutional neural network.

Resumo

O corpo caloso (CC) é a maior estrutura de substância branca do sistema nervoso central e permite comunicação entre os dois hemisférios cerebrais. Esta estrutura é importante uma vez que sua forma e volume estão associados a diversas características da pessoa, doenças neurodegenerativas e não degenerativas e condições clínicas. Normalmente, o CC é estudado usando imagens de ressonância magnética (MRI) onde é segmentado para conseguir extrair informação detalhada e realizar análises posteriores. À medida que a disponibilidade de bancos de dados de MRI cresce e os algoritmos automáticos de segmentação do CC proliferam, verificar o controle de qualidade (QC) é importante para garantir a confiabilidade das segmentações e evitar introduzir erros, que serão propagados ao longo da análise.

Neste trabalho, comparamos dois métodos para fazer QC das segmentações do CC em T_1-MRI sem uso de um *ground-truth*. O primeiro método usa técnicas clássicas de aprendizado de máquina fazendo extração manual de um descritor multi-resolução de forma, junto com um classificador baseado em um comitê de máquinas. O segundo método envolve técnicas de aprendizado profundo usando uma rede neural convolucional (CNN) para extrair descritores da máscara de segmentação e obter informação contextual da imagem. Nos experimentos foram usadas imagens de 907 sujeitos, adquiridas em duas sequências diferentes: 802 de T_1-MRI (incluindo 7 sujeitos com presença de tumor afectando o CC) e 105 de MRI de difusão (procesadas para mapas de anisotropia fraccional). Todas as imagens provêm do hospital das clínicas da Unicamp, a excepção de 247 imagens de T_1-MRI que provêm da iniciativa ABIDE para estudo de autismo.

As duas abordagens propostas para aplicação de QC conseguiram um desempenho similar (área abaixo da curva ROC de 98% aproximadamente) quando treinadas e testadas em T_1-MRI, assim como um tempo de execução idêntico (9 segundos para processar 136 amostras). Adicionalmente, o desempenho dos métodos foi testado em duas bases de dados diferentes à usada para treino: difusão e pacientes com tumor. Em MRI de difusão, a abordagem clássica apresentou melhor desempenho (AUC foi 20% maior), generalizando os padrões aprendidos em T_1. No banco de dados de pacientes com tumor afetando o CC, prevaleceu a CNN com 80% de acurácia, se mostrando mais efetiva dado seu conhecimento do contexto a través da imagem de entrada. Ainda que os dois modelos podem ser usados em MRI, o método profundo é mais versátil, pudendo aprender novas formas e intensidades.

Keywords: corpo caloso; imagens de resonância magnética; controle de qualidade; máquina de vetores suporte; rede neural convolucional.

List of Figures

List of Tables

List of abbreviations and acronyms

ACC	Accuracy
AUC	Area under the curve
CC	Corpus callosum
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
DSC	Dice similarity coefficient
DTI	Diffusion tensor imaging
DWI	Diffusion weighted imaging
FA	Fractional anysotrphy
GM	Grey matter
MRI	Magnetic ressonance imaging
QC	Quality control
QCS	Quality control score
RC	Reverse classifier
RCA	Reverse classification accuracy
RGB	Red, green, blue
RMSE	Root mean square error
ROC	Receiver operating characteristic
SVM	Support vector machine
WM	White matter

Contents

1 Introduction

1.1 Corpus Callosum studies on MRI

The corpus callosum (CC) is the largest white matter (WM) structure in the central nervous system, connecting the hemispheres of the brain and allowing them to communicate (HOFER; FRAHM, 2006). The CC is important in research owing to the correlation between its shape and volume with certain subject characteristics such as gender, aging, and handedness. Some important neurodegenerative diseases, including Alzheimer's and multiple sclerosis, can change the shape and volume of the CC. Besides, other diseases and clinical conditions affect the CC, such as dyslexia, epilepsy, schizophrenia, smoking, obesity, and alcoholism (COVER et al., 2018).

Studies on CC are normally performed using magnetic resonance imaging (MRI), which offers excellent soft tissue contrast and is superior, in general, to other technologies such as radiography and computerized tomography (EDELMAN; WARACH, 1993). While conducting morphological and physiological feature extraction, studies on the CC usually start with CC segmentation (COVER et al., 2018). However, segmenting the CC is particularly challenging because of shape variability among the subjects, the intensity similarity of the CC with neighboring structures (such as the fornix), intensity variability among scanners, the partial volume effect caused by a limited acquisition resolution and artifacts derived from technological limitations such as motion (HE et al., 2007).

With the increasing availability of MRI data and the proliferation of automatic algorithms, segmentation over large datasets has become affordable. Deep learning-based methods are eager for data and their adoption in medical imaging analysis pipelines made populational studies jump from dozens to tenths of thousands of subjects (JR et al., 2008; THOMPSON et al., 2020; KIESOW et al., 2020). In this scenario, manual segmentation is no longer an option because of the high effort involved and time spent. Moreover, a quality control (QC) step is mandatory because segmentation errors can be propagated along the whole pipeline, impairing the final results. There are many automatic and semi-automatic CC segmentation methods, yet none of them are entirely reliable (COVER et al., 2018).

Special attention must be given to images with artifacts (head coverage, radiofrequency noise, signal inhomogeneity, susceptibility, blurring, and ringing) (BACKHAUSEN et al., 2016), newborns (SCHOEMAKER et al., 2016) , young and elder population (WENGER et al., 2014), and ill patients with tumors (GUENETTE et al., 2018). All of these cases present changes in brain morphology or image formation, further limiting the accuracy of automatic segmentation methods and requiring, at least, QC verification.

FreeSurfer[1], for example, is a free and popular and widely used tool for processing and analyzing brain MRI, including segmentation, in both clinical and scientific contexts. However, it makes segmentation errors, especially in subcortical structures such as the amygdala and hippocampus, requiring visual inspection and manual correction of the masks (GRIMM et al., 2015). Moreover, CC FreeSurfer segmentation is performed over the five midline sagittal slices only. However, the CC rarely is confined to these slices. Also, patients with tumors are even more prone to suffer from poor segmentation due to brain symmetry loss. FreeSurfer is atlas-dependant and, for that reason, deficient in these cases. Tailored adjustments can be performed to compensate tumor distortion, but as these adjusts are treated on a case-by-case basis, they are not feasible in large datasets. Likewise, no segmentation algorithm in the literature is 100% effective in segmenting the CC (COVER et al., 2018). All of this makes QC mandatory to assure CC segmentation reliability (GUENETTE et al., 2018).

1.2 Quality control on medical imaging pipelines

Although there is a concern for QC, it is still subjective, prone to errors, and time-consuming, as well as usually conducted in a visual and exploratory manner (REEVES; LIU; XIE, 2016). To determine whether automated subcortical FreeSurfer segmentations are reliable, GUENETTE et al. (2018) visually inspected, and manually corrected the whole T_1 MRI dataset. MAKROPOULOS et al. (2018) proposed a fully automated processing pipeline, including a QC stage, for developing neonatal brain MRI. The QC was performed over cerebrospinal fluid (CSF), WM, and grey matter (GM) segmentation by visually scoring a stratified sampling (10%) of the whole dataset. BACKHAUSEN et al. (2016) introduced a workflow to rate motion artifacts of structural MRI, including a manual verification of segmentation for skull-stripping (removal of non-brain tissue), subcortical/cortical structure borders, and GM using Freeview (FreeSurfer graphical tool). KESHAVAN et al. (2018) developed Mindcontrol, an open-source collaborative web application for brain segmentation QC through a dashboard that allows organizing, exploring, visualizing, annotating, and editing data. Mindcontrol eases the manual quality assurance process, but it does not remove the necessity for manual curation. In summary, QC is applied using visual inspection, in some cases using a graphical tool (Table 1.1). There is no evidence in the literature of the utilization of automatic algorithms for QC in MRI applications.

There is thus a clear need for an automatic QC tool. Some methods, focused on particular applications, have been proposed for perfoming QC. A machine learning approach developed by KLAPWIJK et al. (2019) verified cortical segmentation using supervised random forests, obtaining both high sensitivity and specificity ($AUC = 0.98$).

[1] http://surfer.nmr.mgh.harvard.edu/

Method	Characteristics	Application	Reference
Visual inspection	Verification of the whole dataset	Subcortical Freesurfer segmentation (Head trauma)	GUENETTE et al. (2018)
Visual inspection	Stratified sampling (10%) of the whole dataset	CSF, WM and GM (Neonatal MRI brain)	MAKROPOULOS et al. (2018)
Visual inspection	Graphical tool	Skull-stripping and GM (Rate motion artifacts)	BACKHAUSEN et al. (2016)
Visual inspection	Graphical and collaborative tool	Skull-stripping, WM, GM, CSF (General purpose)	KESHAVAN et al. (2018)

Table 1.1 – MRI studies in the literature applying QC in their processing pipeline

Nevertheless, data was acquired only from one scanner and segmented with FreeSurfer exclusively. The classifier used measures derived from the segmentations rather than masks themselves. Furthermore, no subcortical structures were evaluated.

Outside the medical image context, recent studies showed interest in automatic QC of image segmentation. PENG et al. (2017) presented a framework for evaluating the segmentation quality, composing a reference from multiple labeled segmentations, and using the distance of the composed reference to each new segmentation. In a medical context, ABDALLAH et al. (2016) assessed the low-expertise practitioners' manual segmentations of MRI scans diffused low-grade gliomas using a statistical approach. These two last approaches required several ground-truth to be composed into the reference. SHI et al. (2015) presented an objective measure for visual quality evaluation of an object segmentation using human visual properties as features applied to a common dataset with one foreground object. The quality measured was still subjective and lacked semantic and contextual information.

HUANG; WU; MENG (2016) and SHI; MENG; WU (2017) trained several convolutional neural network (CNN) architectures fusing the segmentation to be evaluated with the original image into the network input. The results were evaluated using the correlation between the output and the existing segmentation evaluation scores. VALINDRIA et al. (2017) and ROBINSON et al. (2017) proposed a QC scheme based on Reverse Classification Accuracy (RCA) with no use of ground-truth. The RCA was tested through three different methods: atlas forests, CNN (DeepMedic (KAMNITSAS et al., 2017)), and multi-atlas propagation. This work was validated over large cardiovascular MRI datasets (ROBINSON et al., 2019), but it would need a re-evaluation for use in other anatomical regions. Moreover, deep learning capabilities were not fully explored. ROBINSON et al. (2018) trained a CNN for predicting the Dice coefficient from 5 regions of cardiovascular MRI. The predicted Dice cannot be used to measure the real Dice, but only for predicting whether segmentation is good or poor given some threshold. ROY et al. (2019) used the QuickNat CNN with a Bayesian extension on four small MRI datasets, measuring the final QC score and the voxel-wise uncertainty map.

Method	Characteristic	Application	Reference
Random forest	Extracted features from segmentations	Cortical Freesurfer segmentation	KLAPWIJK et al. (2019)
Distance	Distance from the segmentation to the labelled references	General RGB images	PENG et al. (2017)
Statistical approach	Statistical analysis over experts segmentations	Diffuse low-grade glioma	ABDALLAH et al. (2016)
Object quality measure	Weighted mean of 4 extracted visual features	General RGB images	SHI et al. (2015)
CNN	Score from mask+image input using CNN's	General RGB images	HUANG; WU; MENG (2016), SHI; MENG; WU (2017)
Reverse classification	Segmentation is used to train 3 RC, which are evaluated on ref. dataset	Cardiac MRI	VALINDRIA et al. (2017)
CNN	CNN is used to predict Dice from mask+image	Cardiac MRI	ROBINSON et al. (2018)
CNN	CNN + Bayesian extension	Brain MRI	ROY et al. (2019)

Table 1.2 – Proposed method to perform QC in images segmentation

In summary, available automatic quality assurance segmentation tools (Table 1.2) present some of the following limitations:

- Require one or several reference segmentations to perform the evaluation. This reference usually is the manual segmentation, which is hard to obtain;
- Take into account only specific descriptors that do not cover all the aspects of the segmentation;
- Use specific or subjective metrics that poorly describe the segmentation quality or accuracy;
- Are established for a specific application or image group, and tested on small datasets, with no guarantee of generalization.

1.3 Objectives

In this work, we pursue a framework for QC of CC segmentations in large datasets without the need for ground truth. For that purpose, we investigated two approaches: classical machine learning and deep learning. The specifics goals of this work are:

- To develop a method for automatic QC of CC segmentations on MRI using classical machine learning techniques;
- To produce a model for automatic QC of CC segmentations on MRI using deep learning techniques;

- To compare the classical approach to the deep learning one, measuring practical issues such as execution time and impact of the dataset size;
- To analyze the extension of both methods to other domains: generalization to other MRI sequences (T_1 and diffusion MRI) and performance on subjects with tumor affecting the CC.

1.4 Main contributions

The main original contributions of this work are:

- Three CC datasets to be made available: 151 T_1-MR images with their respective manual masks, 105 DWI subjects with semi-automatic masks, and 7 T_1-MRI subjects with tumor;
- A multi-resolution shape descriptor for the CC, able to fully characterize the CC shape through a curvature profile;
- A QC method using the proposed CC shape descriptor coupled with an SVM ensemble, able to distinguish correct from incorrect segmentations;
- A QC method on CC segmentations using a simple CNN architecture and contextual information from the image.
- Comparison of both proposed QC methods, the classical machine learning, and the deep learning approaches, in several scenarios: performance and execution time, variation on the dataset size, test on other MR sequence images, and patients with tumor affecting the CC.
- Open source code, available to guarantee reproducibility and make it accessible for researchers interested in perform QC.

1.5 Thesis outline

This thesis was conceived in the format of articles compilation, putting together two published and one submitted articles. Before presenting the articles, a theoretical chapter was included to familiarize the reader with the necessary concepts to understand the posterior chapters. Each article is presented in one independent thesis chapter. It includes a short contextualization of the article in the whole thesis, the presentation of the method and the obtained results, and a final discussion reporting the limitations and how they will be addressed in the next chapter.

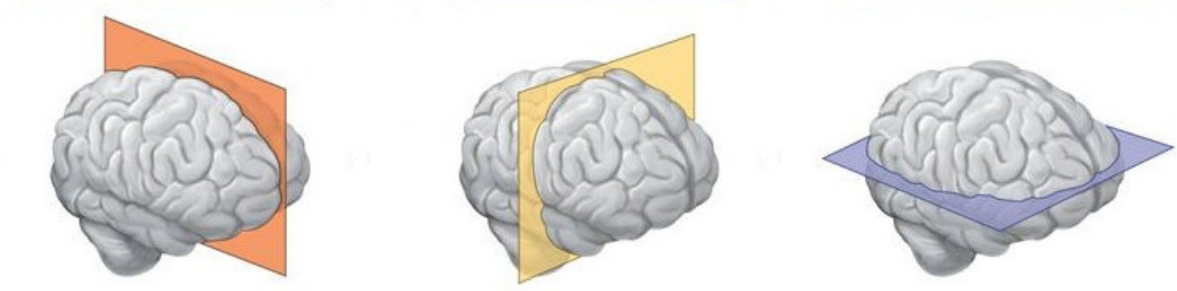
Therefore, the thesis is divided into five chapters: chapter 2 is the theoretical background, including the CC, MRI, and its sequences, and the CC segmentation methods; chapter 3 presents the article entitled **Corpus Callosum Shape Signature for Segmentation Evaluation**, detailing our shape descriptor, called shape signature, and its first usage as a tool to compare CC segmentations from the same subject but in different MRI sequences; chapter 4 introduces the article entitled **A framework for quality control of corpus callosum segmentation in large-scale studies**, where we make use of the shape signature to extract several shape descriptors of the CC at different resolutions and combine them into a classifier mediated by an ensemble to perform QC over the segmentations; chapter 5 presents the article entitled **Automatic quality control on corpus callosum segmentation: Comparing deep and classical machine learning approaches**, describing an additional deep learning approach to measure the quality score of the CC segmentations using a CNN classifier, and comparing both approaches, the classical and the deep, on real domains; and finally, chapter 6 summarizes the results, discusses the findings and future works, and lists the publications and used tools.

2 Theoretical concepts

This chapter addresses the theoretical concepts necessary to understand the next chapters: first, the corpus callosum (CC) being the structure of interest on which we will apply the quality control (QC) methods (Sec. 2.1); then, we will cover magnetic ressoncance imaging (MRI), that is the used technique to visualize and study the CC (Sec. 2.2); in section 2.3 we will study the CC segmentation methods used on MRI. Although MRI has various acquisition sequences, here we will only deal with T_1 and diffusion MRI sequences. Finally, section 2.4 gives an overview of the supervised machine learning approaches as used in this work.

2.1 Corpus callosum

The plane dividing the brain into two symmetrical halves along the inter-hemispheric fissure is called the mid-sagittal plane. In general, any parallel plane to the mid-sagittal one going in the left-right direction is a sagittal slice located on a sagittal plane (Fig. 2.1.(a)). Coronal plane (Fig. 2.1.(b)) is orthogonal to the sagittal plane going from the nape (posterior part) to the noise (anterior part). Axial plane is orthogonal to both, sagittal and coronal planes (Fig. 2.1.(c)), and it goes from the lower part of the structure (inferior) to the upper part (superior) (ENDERLE; BRONZINO, 2012).



(a) (b) (c)

Figure 2.1 – Reference planes used to describe the brain: a) Sagittal plane, b) Coronal plane, c) Axial plane (Source: http://biology-forums.com, 2020)

The CC (tough body in latin) is a structure located underneath the cerebral cortex, is the greatest white matter structure in the central nervous system, with more than 300 millions fibers (HOFER; FRAHM, 2006). The CC connects both brain hemispheres allowing the communication between them. In the sagittal plane, going from the anterior part (A) to the posterior one (P) of the CC, the external portion is known as **genu** (Fig. 2.2(a)) and the lower curve coming out of the genu is called **rostrum** (Fig. 2.2(b)).

Going to the posterior part, in the middle of the CC, it is the **body** (Fig. 2.2(c)) and then the **trunk** (Fig. 2.2(d)). Finally, the posterior edge is known as **splenium** (Fig. 2.2(e).)

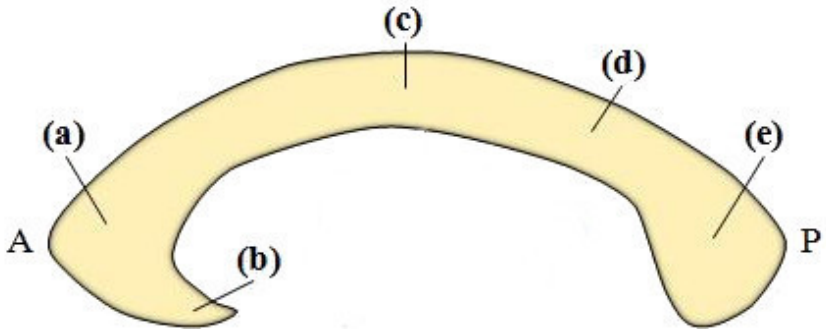


Figure 2.2 – CC parts in the sagittal plane, going in the anterior-posterior direction: a) genu, b) rostrum, c) body, d) trunk, and e) splenium (Source: Segmentação e parcelamento do corpo caloso em imagens de tensor de difusão, 2012 (FREITAS, 2012))

There are several studies on the relationship between the size and shape of the CC and subject characteristics such as sex, age, numerical and mathematical skills, handedness. From a clinical point of view, the CC is affected by illness such as Alzheimer, autism, schizophrenia, dyslexia, epilepsy, multiple sclerosis, depression. Also, the literature relates it with alcoholism, obesity and smoking (COVER et al., 2018).

2.2 Magnetic resonance imaging

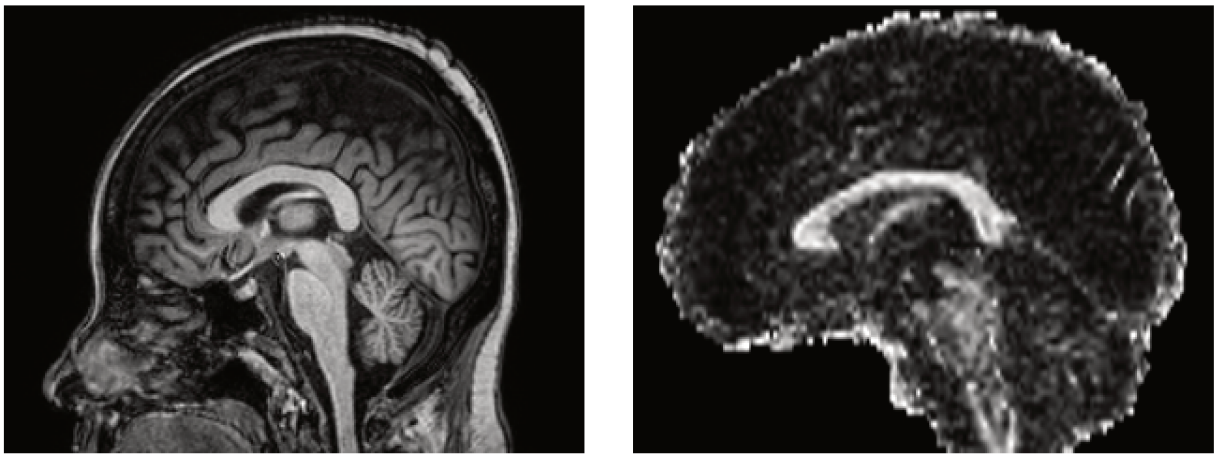
Magnetic resonance imaging (MRI) facilitates the study *in vivo* of the brain structures and their functions. Today, this is the most widely used technique to obtain information about CC as it allows tumor detection, uses non-ionizing radiation, and it is faster and provides better contrast for soft tissue than X-rays and computed tomography (EDELMAN; WARACH, 1993).

Studies in magnetic resonance started formally in 1939, with the technique to measure nuclear magnetic moments (RABI et al., 1939). From there, several studies and experiments were developed around magnetic resonance. However, it was throughout the 1970s that Raymond Damadian, assisted by the Paul Lauterbur's work, developed the theoretical and practical foundations of MRI: created the very first MRI acquisitions, established the time relaxation constants (T_1 e T_2) for detecting cancer tissue and produced the first MRI scanners (DAMADIAN et al., 1976; DAMADIAN, 1971; LAUTERBUR, 1973).

Although the acquisition techniques and the MRI equipment have been improving since then, the principles are the same. The human body is mostly composed of hydrogen atoms that have a nuclear magnetic moment associated, due to their single proton. When a hydrogen atom is positioned in a static magnetic field, its magnetic

moment will align and rotate around the static field, as a gyroscope, at Larmor frequency. To acquire the images, a radio pulse, at Larmor frequency and orthogonal to the static field, is emitted and the hydrogen atom is excited and keeps precesing around the pulse. Then, the radio pulse is turned off, and the magnetic moment returns to its previous state affecting the magnetic field that is detected by the magnetic coil in the gantry. This signal is detected and mapped in position, creating the MRI (ENDERLE; BRONZINO, 2012).

MRI comprises several sequences, also called acquisition protocols, among which we will use in this work: T_1-weighted MRI (Fig. 2.3.(a)) and diffusion weighted MRI. Because diffusion images are complex they are more used in the form of fractional anisotropy (FA) maps (Fig. 2.3.(b)), as we will see in the next sections.



(a) (b)

Figure 2.3 – Imaged brain as appears in: a) T_1-weighted MRI, b) FA map

2.2.1 T_1-weighted MRI

The T_1-weighted MRI (also called T_1-MRI or simply T_1) is acquired mainly from the measurement of the longitudinal relaxation time (T_1), derived of the relaxation signal of the magnetic moment, for a group of atoms located in a specific region (BLOCH, 1946; BLOEMBERGEN; PURCELL; POUND, 1948).

As T_1 is a time constant of exponential growing, tissues with short T_1 are visualized as bright areas, while tissues with longer T_1 are visualized as darker regions (REVETT, 2011). Therefore, the CC can be seen as a bright structure in the center of T_1 images in the sagittal plane, in contrast to the surrounding area (Fig. 2.3.(a)). This sequence is considered structural, because describes the shape and form of the imaged structures.

2.2.2 Diffusion weighted MRI

Diffusion weighted imaging (shortened as DWI, and also called diffusion MRI or simply diffusion) provides contrast based on differences in diffusion in the water molecules within the brain. The very first diffusion sequences were describe in the mid-1960s (STEJSKAL; TANNER, 1965) but only in 1986 were inserted in the medical practice (BIHAN et al., 1986). The diffusion represents the random movement of the molecules (Brownian movement), and depends of several factors such as molecule type, temperature and micro-environment (BAMMER, 2003). In structures with highly oriented fibers, the diffusion along the fibers is greater than the diffusion in any orthogonal direction to them. This diffusion is known as anisotropic, in contrast with isotropic diffusion, where the molecules diffuse equally in all directions (HUISMAN, 2003).

In DWI, the re-alignment of the magnetic moment is affected by the Brownian movement of the molecules, causing signal loss. Therefore, the diffusion can be inferred from this signal loss. In order to describe completely the tissue, gradients of the magnetic field are applied in various directions, thereby achieving one 3D diffusion map for each direction.

This multi-dimensional map is complex, hard to be interpreted and seldom used in the medical practice. The diffusion tensor imaging (DTI) model was introduced to simplify the DWI acquisition (BASSER; MATTIELLO; LEBIHAN, 1994). The diffusion values in all directions are combined, and every voxel (minimum volumetric element in a 3D image) is represented by the second-order tensor $\underline{D}$ (Eq. 2.1) describing the spatial diffusion of the volume (BIHAN et al., 2001; BIHAN et al., 1991).

$$\underline{D} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{2.1}$$

DTI is still a complex model, because of its tensorial elements. A further simplification can be made by deriving a map of anisotropy from the DTI model. Although there are many anisotropy indexes, the most accepted is the FA that can be calculated from the eigenvalues (λ_i) of the tensorial matrix $\underline{D}$ for each voxel using the equation 2.2 (PIERPAOLI; BASSER, 1996).

$$FA = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}{2 \cdot (\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}} \tag{2.2}$$

FA describes indirectly the organization level of the tissue, ranging from 0 for isotropic media (diffusion is the same in all directions) to 1 for completely anisotropic media (diffusion only happens in one specific direction). Because the fibers in the CC are

well oriented, going out of the sagittal plane, this structure has higher anisotropy values than its surrounding structures (Fig. 2.3.(b)) (ABOITIZ et al., 1992).

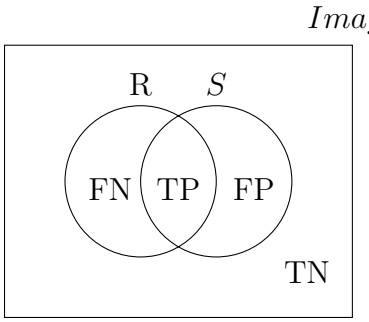
2.3 Corpus callosum segmentation on MRI

The segmentation is the process of partitioning an image into mutually exclusive regions that are spatially adjacent and contain homogeneous pixels (GORDILLO; MONTSENY; SOBREVILLA, 2013). CC segmentation is employed in several applications, including: quantitative study and qualitative visualization, statistical analysis of anatomical variability or deficits, longitudinal monitoring of disease progression or remission, preoperative evaluation and surgical planning (COVER et al., 2018). Although the CC is a noticeable structure, its segmentation is challenging for several reasons: it presents shape variability between subjects, there is intensity variation among different images captured from different MRI scanners using similar sequences, the partial volume effect caused by resolution acquisition, imperfections from equipment pulse profile, and proximity with the fornix (adjacent structure with similar intensity to the CC) (HE et al., 2007). Specially in diffusion MRI, segmentation is hard because of the low resolution of the images.

Manual segmentation methods are commonly used, at small scale, as ground-truth for semi and fully automated algorithms, and in clinical trials, especially where considerable human knowledge and expertise are required to distinguish between brain structures (GORDILLO; MONTSENY; SOBREVILLA, 2013). However, they are not suitable at large scale because demand visual effort, require specialist training and skill, lead to time-consuming processes, and result in both inter- and intra-specialist variability. Semi-automated methods are considered as improved manual implementation, but they cannot fully bridge the manual gaps, and they are still subjected to variability between specialists. Additionally, the intervention of a human operator is often needed to initialize the method, to check the accuracy of the result, or even to manually correct the segmentation result. Fully automated algorithms are efficient and desirable due to their operator-independent nature but they are not as reliable as manual tracing (COVER et al., 2018).

In the literature, there are many methods for segmenting the CC, however none of them is widely used or outweighs all others. Also, there are plenty of metrics intended to quantify the CC segmentation quality (COVER et al., 2018). Among them, Dice similarity coefficient (shortened as DSC, or simply Dice) (DICE, 1945) is a well-established segmentation metric, and the most accepted to evaluate image segmentation performance, that measures the overlap between the mask to be assessed and the reference (commonly the ground-truth) (Fig. 2.4).

In T_1-MRI, there are at least 14 studies among which we can highlight MOGALI et al. (2013) and ADAMSON et al. (2014), that achieve solid results among large dataset



$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.3}$$

Figure 2.4 – Dice calculated as the overlap between segmentation mask to evaluate (S) and a reference (R) from the image representation.

(Dice around 90%). On the other hand, FreeSurfer[1] is a well established brain imaging software package based on atlas model and used for segmenting the structures of the brain. However, its segmentation results are poor, specially in subcortical structures, such as the CC, and in subjects with tumor (GRIMM et al., 2015).

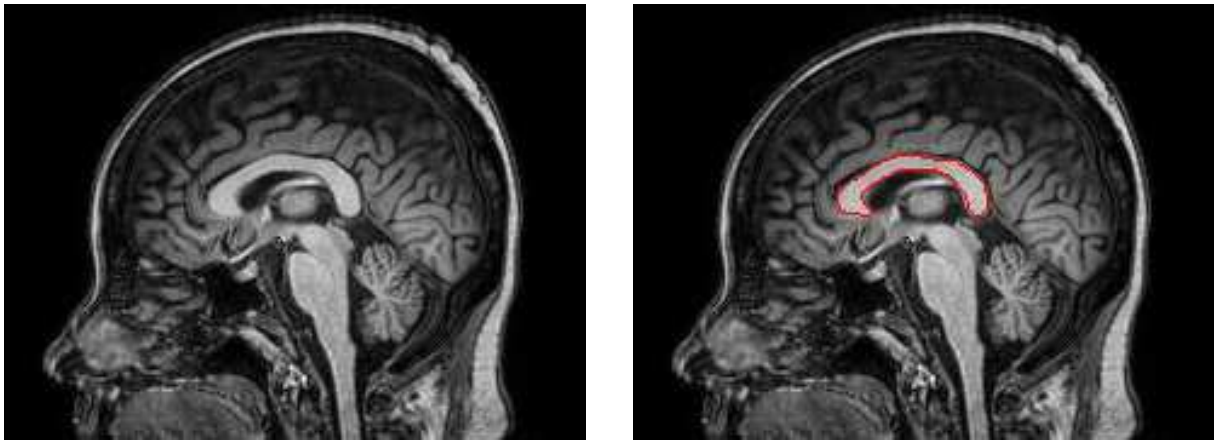
Overall, the CC segmentation is better established in T_1-MRI, when compared with diffusion MRI. In diffusion, among 6 accepted methods, we can highlight NAZEMZADEH et al. (2012) and KONG et al. (2014) that reported a Dice of 96% and 90%, respectively. However, these studies were tested on very small dataset (below 32 subjects) (COVER et al., 2018).

Although there are some proposals to perform the CC segmentation in 3D, most available methods are 2D-based. 2D approach is more used in studies due to its greater usability and quality when compared to 3D segmentations. Also, 2D segmentation gives a suitable CC overview, enough for most practical purposes. When 2D segmentation is adopted, usually the mid-sagittal slice is used, as depicted in the figure 2.5.

2.4 Supervised learning: Classical and Deep

The QC problem can be addressed by classifying a segmentation mask into correct or incorrect class. This classification task can be learned via supervised learning, where a set of rules are created from labelled instances (training set) and applied later to classify new samples (test set) (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). Machines can classify by learning proper features, to guarantee maximum separation between samples from different classes while keeping same-class samples close. In this setting, the classifier can draw a hyperplane to separate the classes. The classification result depends heavily of the features employed by the classifier to represent the samples on the feature space. Further, image applications require the classification function to be sensitive to relevant features, such as shape, while being insensitive to meaningless

[1] http://surfer.nmr.mgh.harvard.edu/



(a) (b)

Figure 2.5 – 2D CC segmentation at the mid-sagittal slice: a) T_1-MRI, b) Highlighted CC segmentation contour.

characteristics, such as position or orientation (DUDA; HART; STORK, 2012; LECUN; BENGIO; HINTON, 2015). In this work, we will compare two distinct machine learning approaches, which will be called **classical** and **deep**.

The **classical** approach extracts handcrafted features, which demands an intensive domain knowledge and engineering skill. Among classical techniques, support vector machine (SVM) is still a state-of-the-art solution for many binary and high-dimensional classification problems. SVM finds the hyperplane with the largest separation margin between positive and negative classes (CORTES; VAPNIK, 1995). The training points lying closest to the decision boundary are called support vectors and determine the separation hyperplane position. When the classes are not linearly separable, a function (kernel) leads the samples to a high-dimensional space, where linear separation is possible, finding a hyperplane with maximal margin of separation between classes (BOSER; GUYON; VAPNIK, 1992). Given a training dataset with N samples ($\{x_i, y_i\}$), where $x_i \in R^n$ are observations with n features and $y_i \in \{0, 1\}$ are their labels, SVM finds the hyperplane with maximal separation margin between classes solving the classification function (Eq. 2.4).

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x_j, x_i) + b \tag{2.4}$$

where $K(x_j, x_i) = \langle \varphi(x_j) \cdot \varphi(x_i) \rangle$ is the kernel that takes the function φ from the input space for the higher dimensionality space ($(R_n \mapsto R_m) : m \geq n$) where classes are linearly separable.

In contrast, **deep** approaches can learn suited features automatically using a general-purpose suervised learning procedure (LECUN; BENGIO; HINTON, 2015). The widespread of deep learning has been possible mainly due to three factors: availability of

computational power; mathematical and empirical techniques for optimizing deep architectures training; and large datasets of labeled information (GOODFELLOW; BENGIO; COURVILLE, 2016). On medical imaging, deep learning has achieved remarkable outcomes in various applications: classification of exams, illness and lesions; detection of tumors, organs, regions and landmarks; segmentation of organs and lesions; registration; big data applications such as content-based image retrieval and combination with reports; and generative models for enhancing, de-noising, normalizing and pattern discovery (LITJENS et al., 2017; LEE et al., 2017).

Among deep learning techniques, convolutional neural networks (CNN) are the most used to deal with images. CNN are feed-forward machine learning models inspired by the visual cortex in the brain. Although there are several CNN architectures, all of them are composed of the same functional blocks made of convolutional and pooling layers. The convolutional ones serve as feature extractor at increasing levels of abstraction, and the pooling ones perform subsampling to reduce computational load and achieve spatial invariance (RAWAT; WANG, 2017).

3 Corpus Callosum shape signature

CC is an elongated bundle of white matter fibers connecting brain cortical hemispheres. Although there is variability in the CC shape among subjects and scanners, in a normal population, the general CC shape can be characterized. In this chapter, a comprehensive shape descriptor is proposed, called shape signature. This name comes from the fact that it is capable of describing the shape with several features, each of them grasping the shape at different resolution and level of detail. In the beginning, the shape signature came from the necessity to describe the shape of a regular CC. However, we soon realized the possibility of applying it to separate populations (e.g., age, sex, illness) and detect incorrect segmentations.

In this chapter, the work entitled **Corpus Callosum Shape Signature for Segmentation Evaluation** is introduced. This work was presented in oral format during the XXVI Brazilian Congress on Biomedical Engineering (CBEB 2018). It was nominated as best regular work in the main category *Cândido Pinto de Melo*, and published in the proceedings of the congress (HERRERA; BENTO; RITTNER, 2019). In this work, the shape signature was applied to the direct evaluation of CC segmentations in diffusion MRI using a ground-truth in T_1-MRI. Because the ground-truth is not present in the same space as the segmentation, the shape signature facilitated the direct evaluation of the segmentation with no additional processes involved, such as registration. Much of the discussion of this work focused on the manual choice of the proper resolution to perform the evaluation.

Using one resolution, the shape signature allowed the evaluation of segmentations from three different methods in diffusion MRI over 145 subjects. However, two critical downsides can be pointed: first, the need for a ground-truth for each subject, making this evaluation method unfeasible in most cases, especially in large dataset pipelines; and second, the manual selection of one resolution to perform the evaluation, discarding the remaining ones, undermines the capabilities of the shape signature to describe the segmentation more richly.

These disadvantages will be approached in future chapters using supervised machine learning. In the meantime, the paper presented in this chapter is essential because it contains the first and full formulation of the shape signature that will be used as the basis for our classical machine learning QC framework (Chpt. 4). Also, the method presented here allows us to evaluate CC segmentations in diffusion MRI where it is harder to obtain the ground-truth, because of the low resolution of the images. It eliminates the need for a registration step that introduces errors into the final evaluation measure.

Corpus Callosum Shape Signature for Segmentation Evaluation

W. G. Herrera, M. Bento, and L. Rittner

Abstract

Corpus callosum is the greatest white matter structure in brain. It is located beneath the cortex and connects both of two hemispheres, making possible their communication. Corpus callosum shape and size are associated with some subject's characteristics such as gender, handedness and age, and alterations in its structure have correlation with some diseases and medical conditions. Diffusion MRI allows a further analysis of corpus callosum structure and functionality by accessing neuronal fibers and tissues microstructure using the water diffusion model. However, the corpus callosum segmentation (required initial step to structural analysis) in diffusion MRI is challenging, since no gold-standard is available. In this work, we propose a segmentation evaluation method that relies on the corpus callosum shape by using its shape signature. We were able to evaluate three different segmentations in diffusion MRI over a 145 subjects' dataset using manual segmentation on T_1 as reference.

Keywords

Corpus callosum • Diffusion MRI • Segmentation • Gold-standard • Shape signature

1 Introduction

The corpus callosum (CC) is the greatest fiber bundle of white matter in the brain allowing communication between left and right brain hemispheres [1]. It is a structure with considerable importance in research, clinical and medical areas since its shape and volume are associated with some subject's characteristics such as gender, handedness and age. CC alterations are related with important diseases and medical conditions such as: Alzheimer, autism, schizophrenia, dyslexia, epilepsy, multiple sclerosis, depression, smoking, alcoholism and obesity [2].

The CC in vivo study is normally performed through magnetic resonance imaging (MRI) due to its better soft-tissue contrast in comparison with other techniques such as radiography and computed tomography [3]. The CC segmentation is a necessary step for any posterior analysis and allows extraction of morphological and physiological characteristics on both, micro and macro levels [4].

The CC segmentation in structural modality T_1-weighted image (T_1) has been widely covered in the literature. However, there are only a few proposed methods in the diffusion space, both in diffusion weighted imaging (DWI) and diffusion tensor imaging (DTI) [2]. Segmentation in diffusion is a challenging task due to: images with low resolution, definition and contrast, CC variability along subjects, intensity variability along scanners, partial volume effect that makes difficult definition of CC borders, proximity and similarity of other structures and thin areas at the CC central zone that causes partition of the structure [5].

In order to evaluate its quality, the segmentation is compared with a reference. An ideal reference is called ground-truth, that is rarely available. A good approximation to the ground-truth is the gold-standard [6]. Normally in neuroimaging, the gold-standard is the manual delineation

W. G. Herrera (✉) · L. Rittner
School of Electrical and Computer Engineering, University of Campinas, Campinas, Brazil
e-mail: w162642@dac.unicamp.br

L. Rittner
e-mail: lrittner@dca.fee.unicamp.br

M. Bento
University of Calgary/Radiology and Clinical Neuroscience, Hotchkiss Brain Institute, Calgary, Canada

R. Costa-Felix et al. (eds.), *XXVI Brazilian Congress on Biomedical Engineering*, IFMBE Proceedings 70/2, https://doi.org/10.1007/978-981-13-2517-5_22

of the structure of interest by a specialist. However, in diffusion, it is hard to obtain the gold-standard because the quality of diffusion images makes difficult manual delineation of the structure.

In the literature, it is possible to find some algorithms for automatic or semi-automatic CC segmentation performing quantitative evaluation [2]. Nazem-Zadeh et al. implemented a 3D CC segmentation over DTI using a level-set method and compared it with a manual segmentation carried out by a specialist [7]. Freitas et al. and Rittner et al. segmented the CC in 2D and 3D using the Watershed transform; both works evaluated the segmentation using a manual segmentation over DTI done by three specialists [8, 9]. Niogi et al. used a threshold method for 2D CC segmentation; the validation was performed by an experienced operator using a semi-automatic software [10]. Kong et al. adopted a graph-based semi-supervised learning model for 3D CC segmentation and used a manual segmentation registered on DTI for evaluation [11]. Garcia et al. implemented a level-set algorithm on DTI for the 3D extraction of the CC; the outcome was validated with tractography directly on DTI [12]. Except for the last work, that followed an unusual approach, all of these works performed validation using a manual segmentation drawn or registered on DTI. Both of these evaluation approaches present some pitfalls: it is challenging to delineate a precise manual segmentation directly on DTI due to low resolution; and the registration process inserts errors during the registration process.

This paper proposes a CC shape signature build by measuring the curvature along CC contour. The proposed signature allows to evaluate and compare segmentations performed in different spaces (diffusion and T_1, for example). Different CC segmentations obtained for a dataset of 145 diffusion images were evaluated against manual segmentations performed on T_1 images, using the proposed shape signature. This study is arranged into five sections as follows: Sect. 2 explains the shape signature and its configuration among extraction, matching and evaluation, Sect. 3 presents the experiments and results regarding our method, Sect. 4 discusses the results and Sect. 5 summarizes the findings obtained in this study.

2 Corpus Callosum Shape Signature

We propose a method based on shape features to directly evaluate CC segmentation in diffusion space, using a gold-standard delineated in T_1, with no registration required.

First, for every segmentation the shape signature is extracted. Then, in order to compare two distinct signatures, matching and evaluation steps are performed.

2.1 Shape Signature Extraction

The proposed signature is a shape descriptor that measures curvature along the segmentation contour. The curvature (k) in one point (p: x_p, y_p) of the contour is given by:

$$k(x_p, y_p) = \arctan\left(\frac{y_{p+r} - y_p}{x_{p+r} - x_p}\right) - \arctan\left(\frac{y_p - y_{p-r}}{x_p - x_{p-r}}\right) \quad (1)$$

where k represents the angle at the p point, between line segments going from (x_p, y_p) to (x_{p+r}, y_{p+r}) and (x_{p-r}, y_{p-r}), the p point is the vertex of the angle and r determines the resolution of the signature. The higher r, the lower the resolution is. The parameter r will be given in percentage of the total length of the contour.

2.2 Shape Signature Matching

Matching of the shape signatures allows fair comparison between signatures from different segmentations because shape signature calculation starts in any arbitrary point along of each segmentation contour. Since the parametric representation of the contour is closed, the signature is periodic. Matching is performed shifting horizontally the signature to be compared maintaining fixed the reference signature. For each position, the distance between signatures is measured using root mean square error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(k_{seg} - k_{ref}\right)^2} \quad (2)$$

where k_{seg} e k_{ref} are the curvature values for the segmentation and the reference, respectively, measured at the p point and averaged over all the P points of the contours. The *RMSE* is measured for all the positions along the contour. The matching position corresponds to the minimum *RMSE*. For every pair of signatures, matching can be performed for different resolutions. Lower resolutions perform well for the matching process because these resolutions hold global information of the segmentation and therefore $r = 0.35$ is a proper value for matching signatures.

2.3 Shape Signature Evaluation

Shape signature evaluation requires always two signatures: the one to be evaluated and the reference one. After matching the signatures, *RMSE* will be used again for quantitative evaluation of the signatures. The higher the *RMSE,* the more distinct the signatures are. Proper resolution

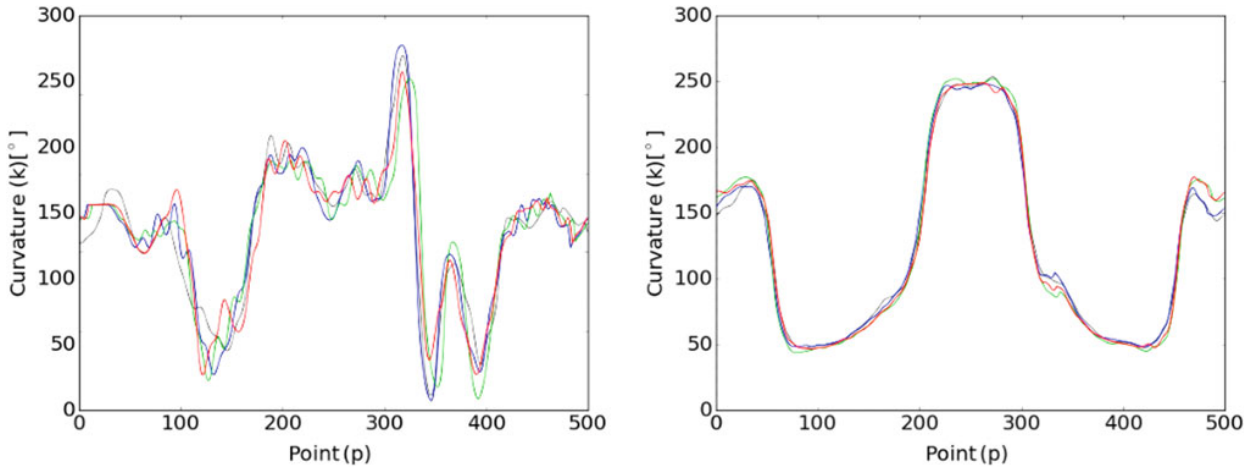


Fig. 1 Shape signatures at different resolutions. (Left) examples of shape signatures at $r = 0.05$. (Right) examples of shape signatures at $r = 0.3$

for segmentation evaluation is a critical hyper-parameter to be chosen. For each segmentation, several signatures can be obtained, one for each resolution ($0.01 < r < 0.49$). Higher resolutions ($0.01 < r < 0.2$) hold details of the contour while lower resolutions ($0.21 < r < 0.49$) describe globally the segmentation shape (see Fig. 1).

In this work, we are interested in compare segmentations of the same subject that tend to differ in finer details and hence, it is expected that higher resolutions will be more suitable to performing evaluation. First, *RMSE* was calculated for a rotated version of the reference, then *RMSE* was calculated for a silver-standard constructed with STAPLE, an independent segmentation tool that computed a probabilistic estimate of the true segmentation [13]. Our goal is to find a resolution that returns a low *RMSE* when a reference is compared with a perturbation of itself (very similar segmentations) while returns a high *RMSE* when a reference is compared with an independent segmentation (distinct segmentations).

Since we are looking for a suitable resolution for evaluation purposes, the *RMSE* for both cases was assessed along all the resolutions using the original reference as basis. The difference between both *RMSE* (rotated reference and silver-standard) was calculated for 50 subjects (see Fig. 2).

As expected, lower resolutions (higher r values) leaded to little *RMSE* differences because signatures, at these resolutions, describe global shape of the segmentation of the same subject neglecting details. Higher resolutions gave larger differences of *RMSE*. Therefore, $r = 0.08$ was the selected resolution for segmentation evaluation because it allowed the higher *RMSE* difference among all the resolutions.

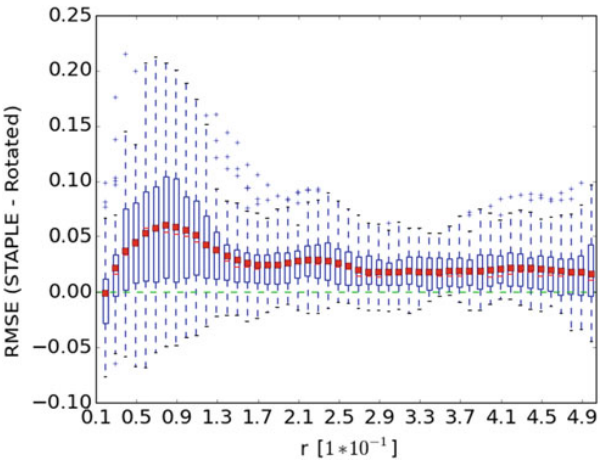


Fig. 2 Mean and standard deviation for the *RMSE* difference along all the resolutions between rotated and silver-standard segmentations

3 Experiments and Results

In order to validate the proposed shape signature, an experiment evaluating three segmentation methods was conducted. MRI from 145 subjects were collected as part of a project approved by the research ethics committee from the School of Medicine at University of Campinas (CEP 920/2007; CAAE: 0669.0.146.000-07). All the participants signed an informed consent form agreeing their participation on the study. All the data was acquired on a Philips scanner Achieva 3T. DWI dataset has a 1×1 mm spatial resolution

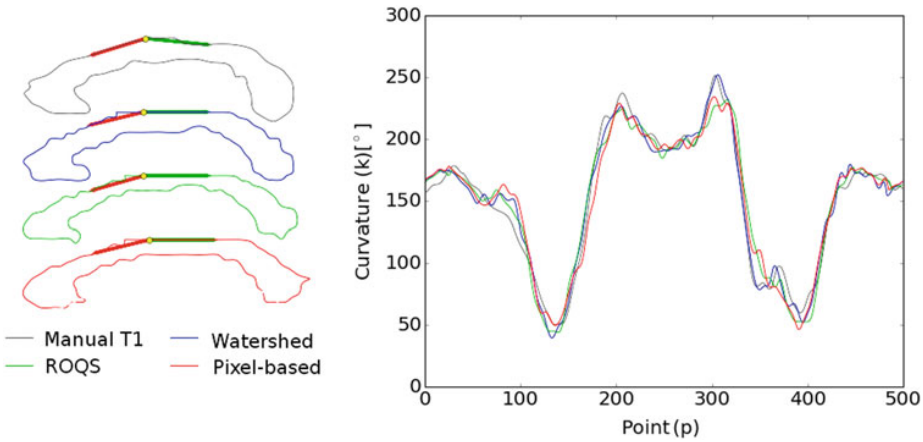


Fig. 3 Shape signatures for three CC segmentations in diffusion and the reference in T_1 (Signatures were matched): (Left) Parametric contour of each segmentation displaying line segments for $r = 0.08$, (right) Shape signatures at $r = 0.08$ associated with CC segmentations

and 2 mm slice thickness in the axial plane, along 32 directions (b-value = 1000 s/mm2, TR = 8.5 s, and TE = 61 ms). T_1 images were acquired in the coronal plane with spatial resolution of 1 × 1 × 1 mm (TR = 7 s and TE = 3.2 ms).

From each subject, the mid-sagittal slice was extracted. The evaluated diffusion segmentations were segmentation based on Watershed [8], Reproducible Objective Quantification Scheme (ROQS) [10] and Pixel-based method [14]. Segmentations were evaluated using as reference manual segmentation, delineated on T_1.

For each subject, the shape signatures associated to the three segmentations were extracted at $r = 0.35$ (for matching) and $r = 0.08$ (for evaluation). Signatures matching process was performed at $r = 0.35$ and the manual segmentation on T_1 was used as reference (see Fig. 3).

For each segmentation in diffusion, it was calculated the *RMSE* along the full dataset at $r = 0.08$ (see Fig. 4).

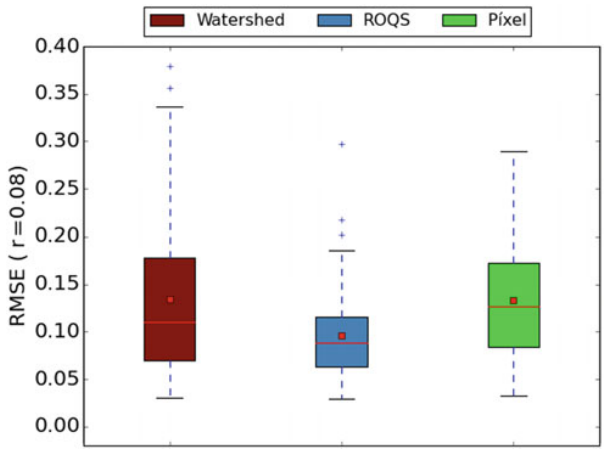


Fig. 4 *RMSE* mean and standard deviation for the three diffusion segmentations along the dataset when used as reference manual segmentation in T_1

4 Discussion

Shape signature is a simple way to compare binary segmentations because computation of the signature itself and the *RMSE* is straightforward. The shape signature is highly related with the resolution that establishes the detail level of the corresponding segmentation. Matching signatures at $r = 0.35$ using *RMSE* presented no errors due to the characteristic shape signature of the CC at this resolution for any subject. For evaluation, a higher resolution ($r = 0.08$) was then used because we were interested in detecting details of different segmentations on the same subject (intra-subject). This resolution ($r = 0.08$) presented a proper tradeoff between noise immunity and details sensitivity. For inter-subject applications, lower resolutions could be more reliable.

Watershed, ROQS and pixel-based segmentations were evaluated in a 145 subjects dataset when compared to manual T_1 segmentation. ROQS was the best method achieving the lowest *RMSE* mean and standard deviation along the dataset; however, it is important to point out that this is a semi-automatic method. On the other hand, the Watershed and the pixel-based methods (both of them fully automatic methods) had similar performances, but their errors were different. While Watershed had more errors (in 20 subjects other regions than the CC were incorrectly segmented), the pixel-based method only failed segmenting the CC for 3 subjects. However, the pixel-based method often segments irregularly the borders due to the pixel-wise approach.

5 Conclusion

In this work, a method for CC segmentation evaluation in diffusion was proposed using the shape signature. This method allowed evaluation of three segmentation methods

using the gold-standard (manual segmentation in T_1) as reference without registering it into diffusion space, since it is time consuming and prone to errors. *RMSE* was used to measure distance between signatures to compare them.

Our proposed shape descriptor may also be used in other applications together with other descriptors and metrics. Some foresee applications are: CC characterization, automatic identification of incorrect segmentations in large datasets, cross-sectional and longitudinal studies regarding CC shape.

Acknowledgements This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES—PROEX 2017). The Program for Partner Graduate Students and National Counsel of Technological and Scientific Development (PEC-PG and CNPq—process 190557/2014-1) and The São Paulo Research Foundation (FAPESP 2013/07559-3).

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Aboitiz, F., Scheibel, A.B., Fisher, R.S., Zaidel, E.: Fiber composition of the human corpus callosum. Brain Res. **598**, 143–153 (1992)
2. Cover, G., Herrera, W., Bento, M., Appenzeller, S., Rittner, L.: Computational methods for corpus callosum segmentation on MRI: a systematic literature review. Comput. Methods Programs Biomed. **15**, 25–35 (2018)
3. Edelman, R., Warach, S.: Magnetic resonance imaging. N. Engl. J. Med. **328**, 708–716 (1993)
4. Rittner, L., Freitas, P., Appenzeller, S., Lotufo, R.: Automatic DTI-based parcellation of the corpus callosum through the watershed transform. Rev. Bras. Eng. Biom. **30**, 132–143 (2014)
5. He, Q., Duan, Y., Miles, J., Takahashi, N.: A context-sensitive active contour for 2D corpus callosum segmentation. Int. J. Biomed. Imaging (2007)
6. Zhang, Y.: A survey on evaluation methods for image segmentation. Pattern Recognit. **29**, 1335–1346 (1996)
7. Nazem-Zadeh, M., Saksena, S., Babajani-Fermi, A.: Segmentation of corpus callosum using diffusion tensor imaging: validation in patients with glioblastoma. BMC Med. Imaging **12**(1) (2012)
8. Freitas, P., Rittner, L., Appenzeller, S., Lotufo, R.: Watershed-based segmentation of the midsagittal section of the corpus callosum in diffusion MRI. In: Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 274–280 (2011)
9. Rittner, L., Campbell, J., Freitas, P., Appenzeller, S., Pike, G., Lotufo, R.: Analysis of scalar maps for the segmentation of the corpus callosum in diffusion tensor fields. J. Math. Imaging Vis. **45**, 214–226 (2013)
10. Niogi, S., Mukherjee, P., McCandliss, B.: Diffusion tensor imaging segmentation of white matter structures using a Reproducible Objective Quantification Scheme (ROQS). NeuroImage **35**, 166–174 (2007)
11. Kong, Y., Wang, D., Shi, L., Hui, S., Chu, W.: Adaptive distance metric learning for diffusion tensor image segmentation. PLoS ONE **9**, 1–11 (2014)
12. Garcia, V., De Jesus, H., Mederos, B.: Analysis of discrepancy metrics used in medical image segmentation. IEEE Lat. Am. Trans. **13**, 235–240 (2015)
13. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging **23**, 903–921 (2004)
14. Herrera, W., Cover, G., Rittner, L.: Pixel-based classification method for corpus callosum segmentation on diffusion-MRI. In: European Congress on Computational Methods in Applied Sciences and Engineering, pp. 217–224 (2017)

4 Quality control of corpus callosum segmentations: classical approach

QC is performed to prevent deficient segmentations from being included in posterior analysis and results. Typical quality evaluation of CC segmentations involves the use of a ground truth per segmentation. In the previous chapter, we used the shape signature to evaluate segmentations in diffusion MRI using the ground truth in T_1-MRI. However, obtaining the ground truth is prone to error and time consuming, making QC unfeasible, especially in large datasets. In contrast, supervised QC schemes can learn particular features to perform segmentation evaluation with no ground truth.

In this chapter, we present the work entitled **A framework for quality control of corpus callosum segmentation in large-scale studies** that was published in the Journal of Neuroscience Methods (HERRERA et al., 2020). In this paper, the shape signature was extracted in 49 different resolutions, from which the more relevant ones were automatically selected using a clustering technique. Then, we put them together in a final SVM ensemble to obtain the final QC measure. The problem was approached with a supervised classifier, trained to distinguish among correct and incorrect segmentations. The final quality measure goes from 0% for correct segmentations to 100% for incorrect segmentations. Because the SVM ensemble learned the proper shape features to perform the classification, there is no need for any ground-truth to use the framework.

The framework was trained and tested exclusively on T_1-MRI, getting an AUC of 98.25% on the test dataset. Because the framework only considers shape features, two critical issues will be analyzed throughout the paper. First, the CC maintains its characteristic shape among different MRI sequences and can be used directly on them. It is important since we can extend the use of the framework to other sequences such as diffusion MRI where manual segmentations are scarce to train the model (low-resolution diffusion images make it more difficult to perform manual segmentation). Second, the framework was trained on a normal population, and therefore it could fail to evaluate abnormal populations such as fetal, newborn, elderly, and tumor patients. These populations are rarely addressed in the literature, but in the medical practice are common and more important to be monitored.

Because the work presented in this chapter only explored T_1-MRI, these two points only will be theoretically addressed in the discussion of the paper. However, they will be revisited in the next chapter, where we will compare this framework with a QC method using deep learning (Chpt. 5) . The comparison will be enriched with experiments in two additional datasets: diffusion MRI and patients with tumor.

Journal of Neuroscience Methods 334 (2020) 108593

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

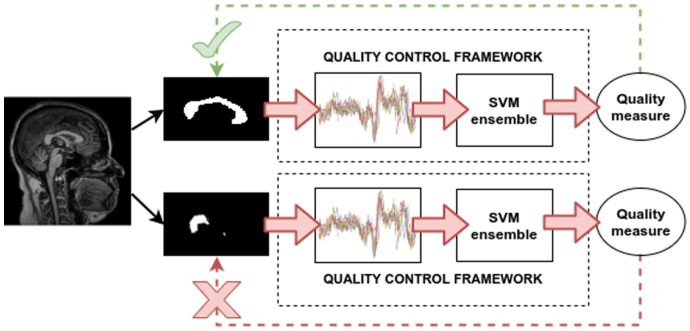
journal homepage: www.elsevier.com/locate/jneumeth

A framework for quality control of corpus callosum segmentation in large-scale studies

William Garcia Herrera[a,*], Mariana Pereira[a], Mariana Bento[b], Aline Tamires Lapa[c], Simone Appenzeller[c], Leticia Rittner[a]

[a] Medical Image Computing Laboratory (MICLab), School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Brazil
[b] Radiology and Clinical Neuroscience, Hotchkiss Brain Institute, University of Calgary, Canada
[c] Rheumatology Department, Faculty of Medical Science, University of Campinas (UNICAMP), Brazil

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Corpus callosum
Segmentation
Quality control
Ensemble
Support vector machine
Magnetic resonance imaging

ABSTRACT

Background: The corpus callosum (CC) is the largest white matter structure in the brain, responsible for the interconnection of the brain hemispheres. Its segmentation is a required preliminary step for any posterior analysis, such as parcellation, registration, and feature extraction. In this context, the quality control (QC) of CC segmentation allows studies on large datasets with no human interaction, and the proper usage of available automated and semi-automated algorithms.
New method: We propose a framework for QC of CC segmentation based on the shape signature, computed at 49 distinct resolutions. At each resolution, a support vector machine (SVM) classifier was trained, generating 49 individual classifiers. Then, a disagreement metric was used to cluster these individual classifiers. The final ensemble was constructed by selecting one representation from each cluster.
Results: The proposed framework achieved an area under the curve (AUC) metric of 98.25% on the test set (207 subjects) employing an ensemble composed of 12 components. This ensemble outperformed all individual classifiers.
Comparison with existing methods: To the best of our knowledge, this is the first approach to assess quality of CC

* Corresponding author.
E-mail address: w162642@dac.unicamp.br (W.G. Herrera).

https://doi.org/10.1016/j.jneumeth.2020.108593
Received 24 September 2019; Received in revised form 9 January 2020; Accepted 9 January 2020
Available online 20 January 2020

W.G. Herrera, et al. Journal of Neuroscience Methods 334 (2020) 108593

segmentations on large datasets without the need for a ground-truth.
Conclusions: The shape descriptor is robust and versatile, describing the segmentation at different resolutions. The selection of classifiers and the disagreement measure lead to an ensemble composed of high-quality and heterogeneous classifiers, ensuring an optimal trade-off between the ensemble size and high AUC.

1. Introduction

The corpus callosum (CC) is the largest white matter structure in the central nervous system, connecting both brain hemispheres and allowing communication between them (Hofer and Frahm, 2006). The importance of CC goes beyond brain interconnection, and its differences in shape and volume have been linked to certain subject characteristics such as sex, aging, and handedness, and, most importantly, to several diseases. Some works have reported association between CC volume with neurodegenerative or inflammatory diseases, for instance, Alzheimer's disease, and multiple sclerosis. In addition, the CC seems to be affected by a number of central nervous system diseases such as dyslexia, epilepsy, schizophrenia, and other common clinical conditions such as smoking, obesity, and alcoholism (Cover et al., 2018). These certain types of conditions might alter CC structure by changing its shape and/or volume.

To verify and track changes in shape and volume and extract morphological and physiological features on the CC, image-based clinical and research studies usually require a preliminary step, the CC segmentation (Gordillo et al., 2013). Segmentation also allows statistical studies along populations, comparison between subjects and individual characterization of the CC. However, CC segmentation through magnetic resonance imaging (MRI) is challenging because of the shape variability among the subjects, the similarity of the CC with neighboring structures such as the fornix, intensity variability among scanners, the partial volume effect caused by a limited acquisition resolution, and artifacts derived from technological limitations such as motion (He et al., 2007).

With the increasing availability of MRI data and the proliferation of automated algorithms, segmentation over large datasets has become affordable (Cover et al., 2018). To avoid errors resulting from the use of poor CC segmentations on the whole analysis pipeline, it is required to practice quality control (QC) over the segmentations. Many QC algorithms rely on the ground-truth, defined as the *correct* segmentation used as reference to evaluate automatic and semi-automatic methods. Commonly, the ground-truth is manually obtained, and that may be a strong limitation, specially in large studies. Besides, the segmentation QC methods are frequently conducted visually, using an exploratory approach, making it subjective, prone to errors, and time-consuming (Reeves et al., 2016).

Although automated methods have been previously proposed for a quality assessment of image segmentation, none of those methods focused on CC. A machine learning approach developed by Klapwijk et al. (2019) verifies cortical segmentation in MRI using supervised random forest algorithm, obtaining both high sensitivity and specificity (AUC = 0.98). In this study, data were acquired only from one scanner and segmented with Freesurfer exclusively. Peng et al. (2017) presented a framework that evaluates the quality of segmentation on generic RGB images. Their assessment is performed by computing the distance between individual segmentations and a reference, composed from multiple labeled segmentations. Abdallah et al. (2016) assessed the manual segmentations of diffused low-grade gliomas on MRI scans using measurements derived from the segmentation and compared them with a reference using a statistical approach. Shi et al. (2015) presented an objective measure for visual quality evaluation of an object segmentation using human visual properties as features applied to generic RGB images.

Among deep learning approaches, Shi et al. (2017) and Huang et al. (2016) trained convolutional neural networks to assess the quality of segmentation using ground-truth segmentations and the corresponding quality score generated from trained datasets. The quality of these methods was measured on generic RGB datasets. Valindria et al. (2017) and Robinson et al. (2017) proposed a QC scheme based on reverse classification accuracy (RCA) in the absence of ground-truth. This last work was validated on a large MRI cardiovascular datasets (Robinson et al., 2019).

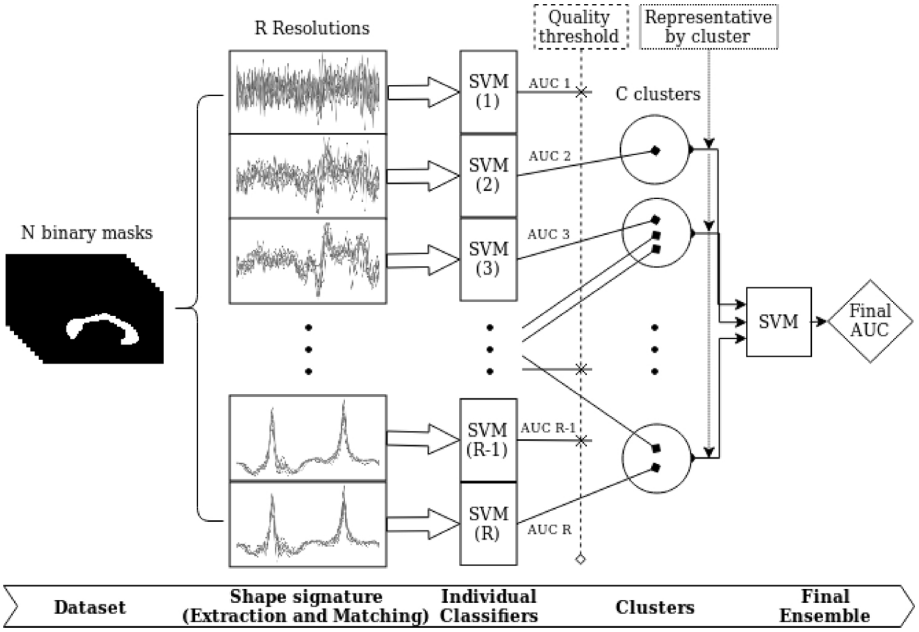


Fig. 1. Diagram of the QC framework: extraction and matching of shape signatures at different resolutions from the dataset composed of binary segmented masks, individual SVM classifiers for each resolution, and the final SVM ensemble composed of high-quality and diverse components. High-quality and diversity was guaranteed by filtering low-quality individual classifiers and grouping by similarity the remaining ones into clusters.

Available automated quality assurance segmentation tools do not fully solve the QC requirements for CC segmentation. In summary, Peng et al. (2017), Abdallah et al. (2016), Shi et al. (2015) require one or several reference segmentations to perform the evaluation, making them impracticable in large dataset analysis; use generic automatic extracted features (Shi et al., 2017; Huang et al., 2016; Valindria et al., 2017; Robinson et al., 2017), or manual extracted features (Klapwijk et al., 2019; Peng et al., 2017; Abdallah et al., 2016; Shi et al., 2015) but not shape; and are established in generic RGB image datasets (Peng et al., 2017; Shi et al., 2015, 2017; Huang et al., 2016) or MRI specific-domain applications (Klapwijk et al., 2019; Abdallah et al., 2016; Valindria et al., 2017; Robinson et al., 2017, 2019), but not CC. Notice only Robinson et al. (2019) tested over large and multiple MRI datasets to consider their work generalizable to similar datasets.

We propose a framework for the automatic QC of CC segmentations in large datasets. Compared to the methods described above, our proposal is the first one to deal with QC of CC segmentation in MRI. It has no need for ground-truth, and it was tested in two large datasets using three different segmentation methods (one manual and two automated). Our framework is based on the CC shape signature and an ensemble of support vector machine (SVM) classifiers. The shape signature is a shape descriptor extracted by measuring the curvature along the segmentation contour and offering the shape characterization of the CC at different resolutions.

This paper is divided into four sections: Section 2 describes the components of the proposed framework namely: shape signature extraction and matching, SVM individual classifiers and the ensemble that outputs the final quality measure; Section 3 describes the experiments, including the used datasets, the chosen hyper-parameters, the final results over the test set and two additional experiments to evaluate the framework reliability; and Section 4 discusses the results highlighting the importance of our framework, the used criteria to construct the ensemble, the final quality measure, and the usage and limitations of our method in real applications. Finally, Section 5 provides some concluding remarks summarizing the current study and presenting possible extensions of our work.

2. Methods

Our framework allows an automatic assessment of the quality of CC segmentations. CC has large variability among the subjects, and its shape maintains a characteristic pattern that can be used as a descriptor for evaluating the quality of the segmentation. The use of a shape descriptor as an attribute to describe the segmentation quality has two main advantages; namely, there is no need for a ground-truth for each segmentation, and the signature can capture the segmentation shape at various levels of detail (resolutions), making the classification highly customizable.

In this work, we make use of supervised machine-learning technique to learn the characteristic patterns of the CC shape and to choose the most relevant signature resolutions to perform the classification task. The supervised classifier learns to distinguish between correct and incorrect segmentations. Then, it assigns to a new segmentation a probability, which ranges between 0% for completely correct segmentation, and 100% for completely incorrect segmentation. This probability is the quality score.

The proposed framework consists of three main components to achieve the QC of CC segmentation (Fig. 1): Sections 2.1 and 2.2 describe the extraction and matching of the shape signatures from the CC segmentation masks, Section 2.3 describes the supervised individual classifier using an SVM that receives the segmentation shape signature as input and outputs the probabilities for each resolution, and finally Section 2.4 details the combination of individual SVM classifiers into an ensemble that takes the individual probabilities and combines them in a final agreed probability: the quality score.

2.1. Shape signature extraction

Our method is based on the shape signature, name given to a shape descriptor that measures the curvature along the CC segmentation contour at several resolutions (Herrera et al., 2019). Shape descriptors representing contours at different resolutions have already been used for content-based image retrieval application or contour description. Adamek and O'Connor (2004) extracted the multi-resolution descriptor by evolving the segmentation contour and representing it through a 2D matrix; Mokhtarian and Mackworth (1992) constructed a multi-resolution descriptor by convolving the parametric representation of the contour with Gaussian functions of different parameter values; Jomma and Hussein (2016) proposed a multi-resolution shape descriptor by measuring the distance from the segmentation contour points to each viewing point on several circular orbits positioned at the segmentation centroid. Only Mokhtarian and Mackworth (1992) descriptor is similar to ours, although the curvature is computed differently. None of these methods use the multi-resolution shape descriptor for segmentation assessment, medical imaging or QC. Furthermore, none of these methods discusses selection of the proper resolutions or even the use of the shape descriptor in a machine learning ensemble classifier.

For our shape signature, the curvature k in one point p (x_p, y_p) is given by:

$$k(x_p, y_p) = \arctan\left(\frac{y_{p+\text{ext}_y} - y_p}{x_{p+\text{ext}_x} - x_p}\right) - \arctan\left(\frac{y_p - y_{p-\text{ext}_y}}{x_p - x_{p-\text{ext}_x}}\right) \quad (1)$$

where k represents the curvature (angle) at point p, between line segments $(x_{p-\text{ext}_x}, y_{p-\text{ext}_y})(x_p, y_p)$ and $(x_p, y_p)(x_{p+\text{ext}_x}, y_{p+\text{ext}_y})$, and ext ($\overrightarrow{\text{ext}} = \overrightarrow{\text{ext}}_x + \overrightarrow{\text{ext}}_y$) determines the resolution of the shape descriptor. Note that the extension (ext) is the opposite to the resolution: the greater the ext the smaller the resolution. As ext increases, the curvature (k) loses in detail, in other words, its resolution decreases.

The curvature computation is easier through a parametric spline representation of the contour. In this case, ext is given as a percentage of the total parametric length of the contour. ext is only measured between 0 and 0.5 ($0 < ext < 0.5$) because, for a greater ext ($0.5 < ext < 1$), the shape signature is mirrored (Fig. 2).

Signature is extracted by calculating curvature k (Eq. (1)) in every equidistant p point for the total number of points P, along the closed contour. This signature is calculated for several resolutions (varying ext) gathering shape representation in various levels of detail. Both P and ext, are chosen as part of the framework arrangement.

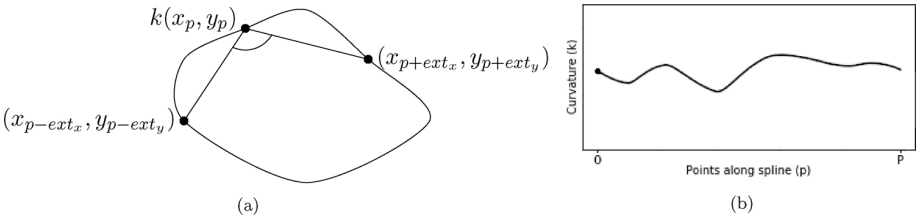


Fig. 2. Signature extraction: (a) Curvature measure k for a contour at point p (x_p, y_p). The curvature is measured as the angle between lines going from point p to point (x_{p-ext}, y_{p-ext}), and point p to point (x_{p+ext}, y_{p+ext}), and (b) Shape signature when k is measured along the contour.

2.2. Shape signature matching

Shape signature matching is a process in which two or more shape signatures are shifted to the same relative position, thereby allowing a comparison to be conducted. By finding the optimal matching position, the reference shape signature is fixed, and the remaining ones are shifted for all points p (recalling that the curvature descriptor is periodic because the contours are closed). The distance between each remaining signature and the reference is measured using the root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(k_{\mathrm{ap}} - k_{\mathrm{bp}})^2} \tag{2}$$

where k_{ap} and k_{bp} represent the curvature for segmentations a and b, respectively, measured at point p. P is the total number of points where the curvature is calculated. The matching point is where the RMSE is the smallest (Chao and Chien, 2010). Matching at resolutions between 0.25 and 0.45 ($0.25 < ext < 0.45$) is performed well owing to the fact that, at these resolutions, the signature represents the global segmentation shape and is not influenced by subtle shape variations. On the other hand, at higher resolutions ($0.01 < ext < 0.05$) the signature is prone to noise, and therefore such resolutions are inappropriate for matching. The signature shift applied during the matching process is repeated for all resolutions of each segmentation, thereby ensuring a match for every resolution.

2.3. Supervised classification using support vector machines (SVM)

In our framework, the input is a binary segmentation mask, which can be obtained through the execution of any automated, semi-automated or manual segmentation method. Afterward, each segmentation is assigned a probability of belonging to one of two classes: correct (negative) or incorrect (positive). The higher the probability is, the more likely the segmentation is incorrect. This probability is the quality measure of our framework. By applying a threshold (decision threshold) to the quality measure it is possible to classify a new segmentation between correct or incorrect. Therefore, segmentations above the decision threshold are classified as incorrect and below as correct.

The classification task can be learned using a supervised machine learning scheme. A supervised classifier creates a set of rules from labelled instances (training set) and uses them to classify new instances (test set) in a process called generalization (Kotsiantis et al., 2007). A validation set is used to avoid over-training that happens when the classifier learns specific rules only applicable to the training set but not generalized to the test set.

Among supervised techniques, SVM is still a state-of-the-art solution for many binary and high-dimensional classification problems. SVM is a type of supervised machine learning technique that will be used for both, the individual classifiers and the ensemble. SVM finds the hyperplane with the largest separation margin between positive and negative classes (Cortes and Vapnik, 1995). The training points lying closest to the decision boundary are called support vectors and determine the separation hyperplane position. When the classes are not linearly separable, a function (kernel) is used to map the points to a higher dimension space, where the classes are linearly separable (Burges, 1998). As mentioned, the classifier outputs the probabilities for each class and the decision threshold must be applied to assign the instance to a certain class.

The selection of the proper way to evaluate a classifier is not an easy task because there are many available metrics and their selection depends on the final application (Sokolova and Lapalme, 2009). The receiver operating characteristic (ROC) curve is a graphical tool for evaluating a classifier by considerating the true positive rate (TPR) and the true negative rate (TNR) for the full range of decision threshold (Hanley and McNeil, 1982) (Fig. 3).

Every point in the curve is given by a decision threshold, leading to a trade-off between TPR and TNR maximization. We are not interested in evaluating the classifier based on a particular decision threshold but assessing the performance over its entire operating range. Area under the curve (AUC) is a well known and versatile metric for classification tasks, whose value does not depend on the selected decision threshold. A real classifier falls between the "random guessing" classifier (AUC = 50%) and the ideal classifier (AUC = 100%). By maximizing the AUC, the classifier will have good performance at any decision threshold (Hajian-Tilaki, 2013). Therefore, AUC is used for quality assessment of the classifiers by allowing the filtering of low-quality individual classifiers.

We can classify segmentations between correct or incorrect by defining a decision threshold using the $F1_{score}$ that takes both, TPR and TNR, into account. In the literature, the $F1_{score}$ is presented as the weighted average of precision (agreement of the real positive classes with those of the classifier) and recall (effectiveness of a classifier to identify positive instances) (Powers, 2011). It can be noticed that, as the decision threshold increases, precision increases (TPR increases), but recall decreases (TNR decreases). A good choice is selecting the decision threshold associated with the maximum $F1_{score}$, thus obtaining the best trade-off between recall and precision.

$$F1_{\mathrm{score}} = \frac{2*(\mathrm{precision}*\mathrm{recall})}{(\mathrm{precision} + \mathrm{recall})} \tag{3}$$

2.4. Ensemble of individual classifiers

For each binary segmentation, one signature for each resolution is extracted (as described in Section 2.1). Different resolutions are chosen to represent the shape with varying levels of detail. After obtaining the signatures for the entire dataset, they are matched (as detailed in Section 2.2). Thus, a 3D feature matrix of size [N, R, P] is assembled, where N is the number of samples, R is the number of resolutions extracted per segmentation, and P is the number of points where the curvature is calculated.

It is difficult to determine which resolutions lead to an optimal separation between correct and incorrect segmentations. For determining significant errors in the segmentation, the global shape is essential; although with this representation, small errors can be neglected. Mixing all R resolutions in a unique classifier does not guarantee good results because each resolution has a different discriminatory power, and there is a lot of redundancy among the data. For this reason, R individual SVM classifiers, one for each resolution, are trained and evaluated (as described in Section 2.3).

After training, the individual SVM classifiers give distinct

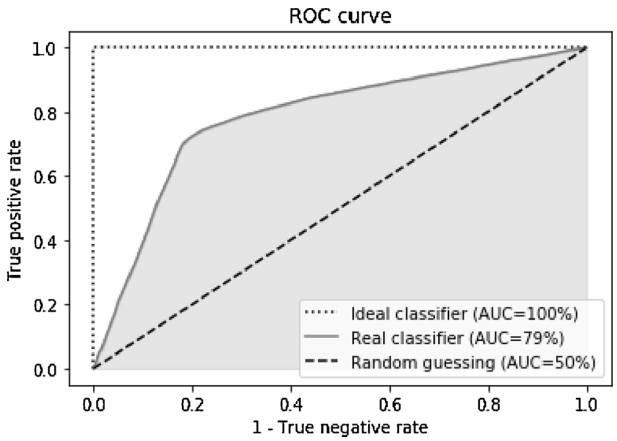


Fig. 3. ROC curve depicting the ideal classifier, the random guessing case and an example of real classifier with associated AUC.

probability results. Some of them will be selected and passed as inputs to the ensemble. Bagging and boosting are popular techniques for constructing ensembles based on their simplicity and effectiveness; as shown in Faria et al. (2012), a bagging SVM ensemble presents superior results. The ensemble of classifiers aims to surpass the performance of individual classifiers, easing the agreement of several classifiers and increasing the generalization capability of the framework. The selected classifiers whose outputs are passed to the ensemble are called ensemble components.

An effective ensemble uses the minimum number of high-quality components trying to combine heterogeneous classifiers complementing each other (Tulyakov et al., 2008). So, the aim is first to select **high-quality** classifiers, using the quality threshold[1] (based on the AUC). Therefore, AUC is used for filtering low-quality individual classifiers, selecting only the highest AUC classifiers on the validation set (individual classifiers above the quality threshold). Then, from the selected ones, we find the most **diverse** ones, based on the disagreement measure. There are many different useful measurements used to select heterogeneous classifiers (Kuncheva and Whitaker, 2003). These measures are based on the agreement matrix (Table 1) where True corresponds to cases correctly predicted by the classifier, and False corresponds to cases incorrectly predicted.

The disagreement measure *Dis* (Eq. (4)) defines the proportion of the number of observations in which one classifier correctly predicts an outcome, and the other incorrectly predicts it, and vice-versa, to the total number of observations.

$$\mathrm{Dis}_{c_i,c_k} = \frac{b + c}{a + b + c + d} \tag{4}$$

Dis is a pair-wise metric that focuses on identifying classifiers that accomplish different types of False results. A matrix can be obtained depicting the disagreement between all pairs of classifiers. This matrix is used as a distance matrix where highly discrepant classifiers have higher values and are represented farther away from each other. Therefore, a clustering method is used to group classifiers, given the distance between them. A recursive algorithm forms initial clusters and merges the pair of clusters that minimally increases the *Dis* distance. The number of clusters becomes a hyper-parameter that must be defined.

From each cluster, the classifier with the highest AUC is chosen. The selected individual classifiers are used to compose the final ensemble. The ensemble is formed by an SVM whose inputs are the probability outputs of the individual classifiers and returns the probability of a given segmentation being incorrect.

3. Results

In order to validate the proposed framework, we first trained individual classifiers and then an ensemble to classify new segmentations. AUC was used to evaluate the whole framework. To obtain the final classification into one of the following two classes: *Correct* and *Incorrect*, decision threshold was experimentally determined using $F1_{score}$. Because both of the classification steps, namely, individual classifiers and ensemble, use a supervised scheme, the segmentations were labeled as *Correct* (Fig. 4a) and *Incorrect* (Fig. 4b and c) using a manual approach. Signatures were extracted for both of them (Fig. 4d and e).

3.1. Data preparation

To increase the generalization, T1-MR images were obtained from two studies with three different segmentation methods. From the 548 T1 images acquired at the University of Campinas, 397 were segmented with Freesurfer v5.3.0, which is freely available for download online[2] and the remaining 151 were manually segmented by a specialist. On the other hand, 247 subjects from ABIDE database (Hiess et al., 2015) were automatically segmented and manually corrected in minor details by Ardekani (2013). All data used in preparation for this article were approved by the local ethical committee and fully anonymized. All of the participants were duly informed, and all of them signed a consent form agreeing to participate in the studies. Due to property rights, the data cannot be shared with the community. The manual labelling of the segmentations was performed by the authors in a one-to-one visual basis. As it will be pointed out later in this section, for precaution, segmentations whose class was not clear were discarded.

Table 1
Agreement matrix for classifiers c_i and c_k, where a represents cases in which both the classifiers correctly predicted the result; b represents cases in which c_i correctly predicted, and c_k incorrectly predicted the result; c represents cases in which c_i incorrectly predicted, and c_k correctly predicted the result; and finally, d represents cases in which both the classifiers incorrectly predicted the result.

	True c_k	False c_k
True c_i	a	b
False c_i	c	d

The experiments were conducted using a Python/Numpy environment (Oliphant, 2006) along with Scikit-learn (Pedregosa et al., 2011) for machine-learning implementations. Both, the source code and the saved trained model can be found on GitHub.[3]

From each T1-MRI volume, only the mid-sagittal slice was employed using the acpcdetect tool from ART toolbox (Ardekani et al., 1997). After segmentation and mid-sagittal selection, some images were discarded due to registration problems, bad selection of the mid-sagittal slice or it was not clear which class the segmentation belonged to (subtle errors in the segmentation). Therefore, the final dataset was composed of 688 segmentations, distributed into 287 incorrect segmentations (42%) and 401 correct segmentations (58%).

The shape signatures were extracted from the final segmentation dataset by firstly obtaining the contour (S), by applying a logical XOR, pixel-wise, between the original segmentation (G) and its eroded version using a structuring element (e) of size 1 (Eq. (5)).

$$S = \mathrm{XOR}(G, G \ominus e) \tag{5}$$

Then, for each contour, a spline of degree $g = 5$ and smoothness of 700 was calculated. The shape signatures were extracted by measuring the curvature of 500 points ($P = 500$) along the spline for 49 resolutions ($R = 49$): from 0.01 to 0.49 with steps of 0.01 (Fig. 5). Finally, all signatures were matched using $ext = 0.35$, resulting in a 3D feature matrix of [688, 49, 500].

3.2. Individual classifiers

Prior to the classification, the feature matrix was randomly divided into three sets: $train_{ind}$, $train_{ens}$, and $test_{ens}$ (Table 2). Individual SVM classifiers were trained using the $train_{ind}$ set. A grid search to adjust the SVM was applied along with the kernel type (kernel, {rbf, linear, polynomial}) and penalty parameter (C, {0.1,1,10,20,50,100}) using a cross-validation strategy. For illustration purposes, the final configuration of each SVM-resolution classifier and one shape signature example based on the resolution are presented in Table A.1 in Appendix A.

[1] The reader should not confuse the quality threshold with the decision threshold. The quality threshold is applied over the individual classifiers to select only high-quality components while the decision threshold is applied on the final output probability to classify the segmentation as correct or incorrect.

[2] http://surfer.nmr.mgh.harvard.edu/.

[3] https://github.com/wilomaku/CC_seg_clas.

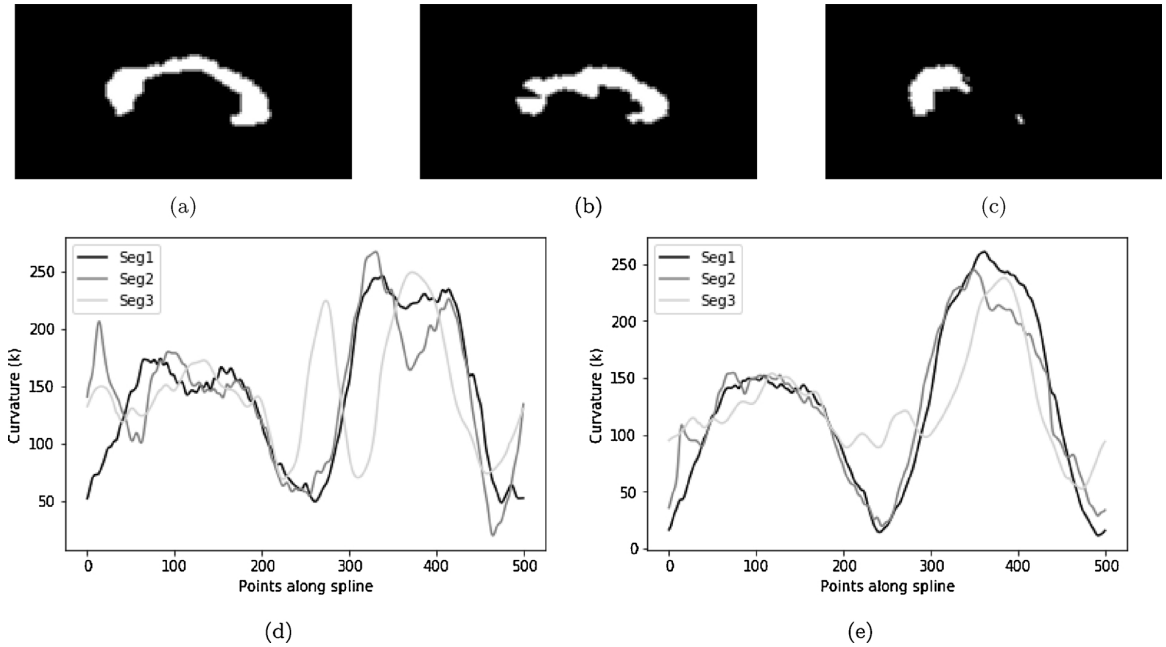


Fig. 4. Example of three segmentations: (a) Seg1 – correct segmentation, (b) Seg2 – incorrect segmentation, and (c) Seg3 – incorrect segmentation with their associated signatures at resolutions: (d) *res* = 0.10, and (e) *res* = 0.15.

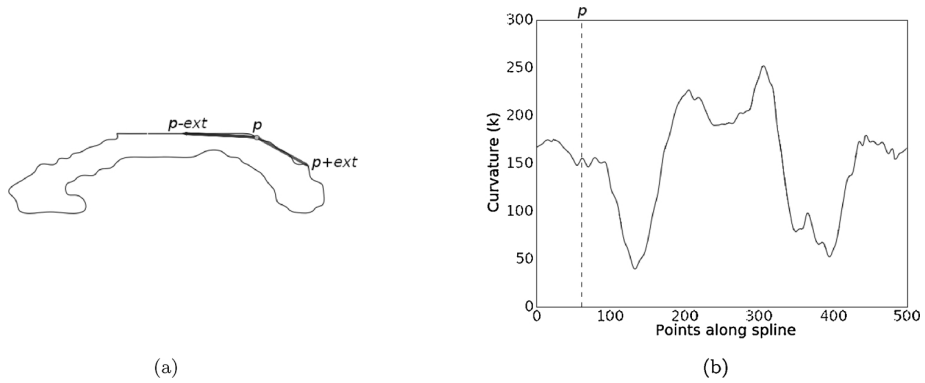


Fig. 5. Extraction of shape signature: (a) Spline depicting one example point p where the curvature is measured between the anterior ($p - ext$) and posterior ($p + ext$) points, and (b) Shape signature measured along 500 contour points obtained at $ext = 0.1$.

Table 2
Final dataset distribution for training and test sets both the individual classifiers and the ensemble.

Set name	Samples	Proportion	Purpose
$train_{ind}$	240	35%	Training and validation of the individual classifiers
$train_{ens}$	241	35%	Training and validation of the ensemble
$test_{ens}$	207	30%	Test of the ensemble

3.3. Ensemble of classifiers

To build the ensemble, two criteria were met: quality and diversity. Regarding quality, AUC was used to measure classification performance of the individual classifiers. Individual classifiers below the quality threshold ($AUC < 0.9$) were rejected: 8 classifiers were discarded, and 41 remained (Fig. 6a). Afterward, using the $train_{ens}$ set, a disagreement matrix (Fig. 6b) was constructed in which rows and columns represent the individual classifiers and each cell represents the disagreement measure (Eq. (4)) between every pair of classifiers. Lighter values denote lower disagreement (less diversity) and darker values mean higher disagreement (more diversity).

The disagreement matrix was used as a distance matrix to determine the optimal ensemble composition by grouping nearby classifiers (lower disagreement). First, the agglomerative clustering technique was applied to identify classifier clusters using the disagreement distance. Further, for each cluster, only one individual classifier was chosen (cluster representative) to construct the ensemble by electing the classifier with a minor intra-cluster distance. Therefore, the number of clusters is the same as the ensemble size because only one classifier was chosen from each cluster. An extensive search over all the possible number of clusters was conducted from 1 to 41, as it is not possible to define the number of clusters a priori.

Forty-one ensembles were trained in the $train_{ens}$ set, one for each ensemble size, along with the same grid search used for individual classifiers. The AUC for the $test_{ens}$ set was obtained (Fig. 7a). The optimal ensemble size was chosen as the minimum size that achieved the highest AUC, obtained along all the possible ensembles. In this case, the optimal ensemble size was 12.

The optimal ensemble achieved an AUC of 98.25% (Fig. 7)b. The

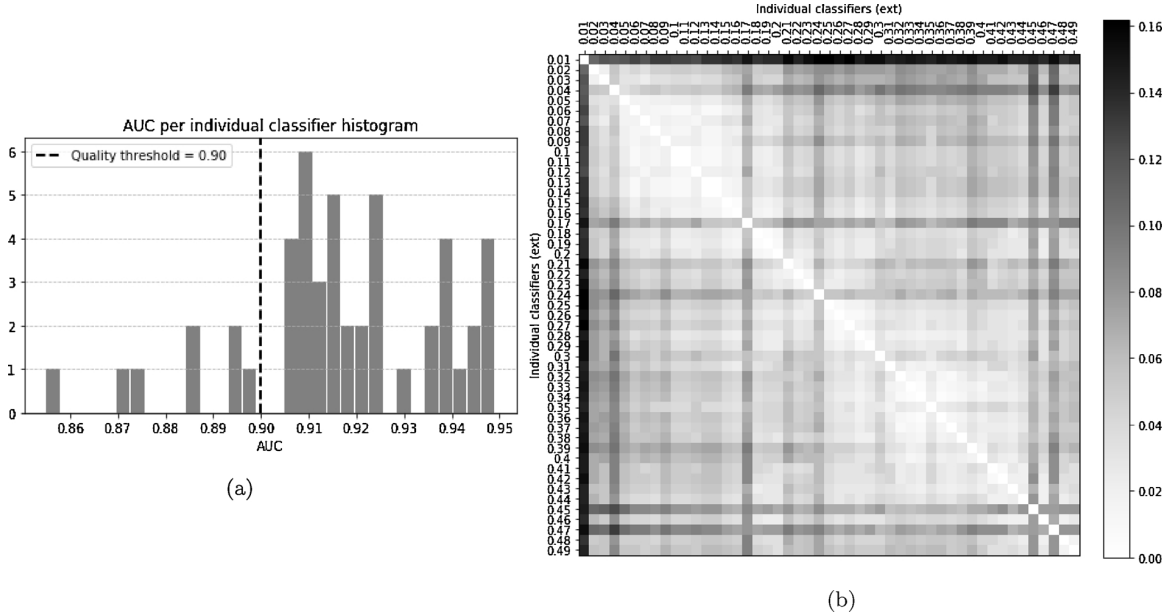


Fig. 6. Criteria for ensemble construction: (a) Histogram for the AUC of individual classifiers where quality threshold was applied (classifiers with $AUC < 0.9$ were rejected) and (b) Disagreement matrix representing the diversity between every pair of classifiers. Each cell in the matrix is the disagreement measure *Dis* (Eq. (4)) between the row and the column classifiers.

resolutions (ext), associated to the individuals classifiers, selected for conforming the ensemble were *0.03*, *0.04*, *0.05*, *0.16*, *0.23*, *0.28*, *0.31*, *0.34*, *0.39*, *0.44*, *0.46*, *0.48*. To classify segmentations as correct or incorrect, the decision threshold, defined at the maximum $F1_{score}$, was 31%. Applying this decision threshold to the output probability, the ensemble presented 9 miss-classifications (accuracy = 95.65%) (Fig. 8).

3.4. Additional experiments

Two additional experiments were performed to evaluate the reliability of our framework. In the first experiment, we tested the **importance of diversity** by comparing the optimal ensemble with a high-quality ensemble with no diversity. In the second experiment, the **sensibility to hyper-parameters** was tested by varying them and checking the output.

Importance of diversity: As previously mentioned, diversity is crucial for the selection of suitable ensembles. A new ensemble was mounted with the top 12 high-quality individual classifiers, whose resolutions (ext) were: *0.05*, *0.06*, *0.07*, *0.08*, *0.09*, *0.10*, *0.11*, *0.12*, *0.13*, *0.14*, *0.15*, *0.16*. This ensemble had the same size as the optimal one, however it had poor diversity because the associated resolutions were close to each other. The 12 high-quality ensemble achieved 96.81% of AUC against 98.25% achieved by the optimal ensemble, in the $test_{ens}$ set (Fig. 9).

Sensibility to hyper-parameters: Although the ensemble has the advantage of automated selection of the proper resolutions to achieve the classification task, its disadvantage is the need to adjust some hyper-parameters: the number of extracted resolutions (R), the quality threshold applied to AUC to filter low-quality individual classifiers, the dissimilarity measure, and the criteria to select representatives by cluster. This experiment aimed to test the framework robustness to variation of these hyper-parameters with regards to the original selected values. The initial hyper-parameter values were modified one by one, while maintaining the remaining ones invariant, obtaining the final AUC (Table 3). These values were compared with the AUC for the optimal ensemble with its original values (last row of the table).

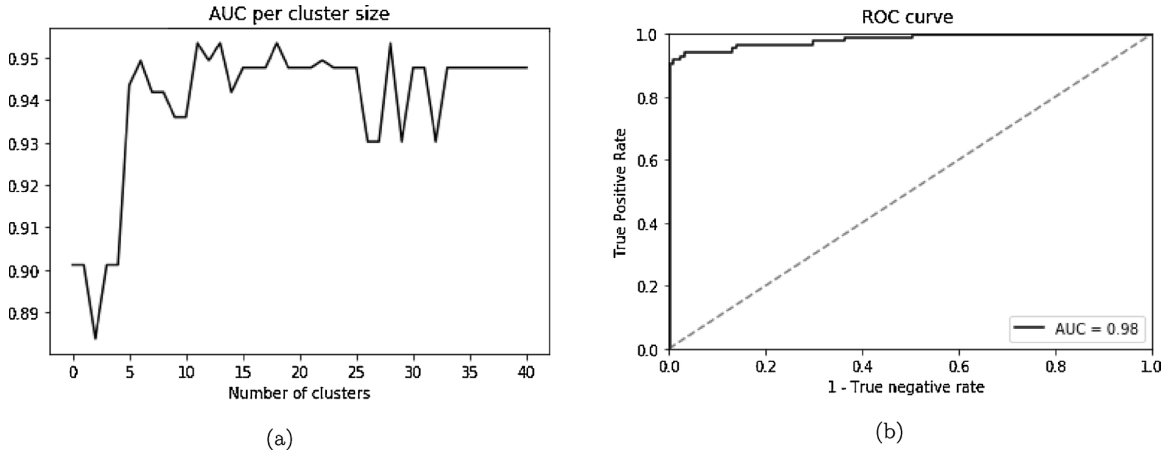


Fig. 7. Optimal ensemble: (a) Extensive search over all the possible number of clusters (ensemble size) depicting AUC for each ensemble, and (b) Final ROC and AUC for the optimal ensemble.

Fig. 8. Confusion matrix and false outcomes cases for the ensemble presented at *threshold* = 31%: (a) confusion matrix, (b) false negative example with p = 6.8%, (c) false positive example with p = 41.6%.

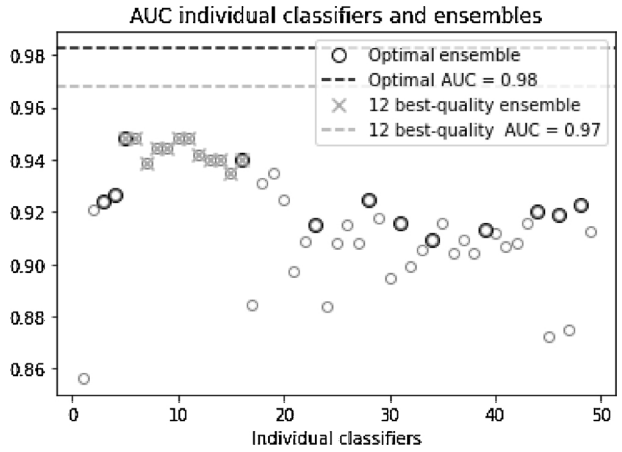


Fig. 9. Comparison of the AUC from optimal ensemble, best-quality ensemble and individual classifiers on the $test_{ens}$ set. Each points is the AUC associated to one individual classifier along the 49 resolutions (0.01 to 0.49). The components of both the two ensembles are depicted: optimal in bold and best-quality like cross. The optimal ensemble AUC is superior to the best-quality one's and both of them are superior to AUC of any individual classifier.

4. Discussion

Quality control of segmentations is an essential step in MRI analysis. Traditionally, QC of segmentations is performed manually, but this is becoming unfeasible with the rise of the large datasets of today (Reeves et al., 2016). Other authors proposed methods to assess QC of segmentations in MRI using specific-domain features: Klapwijk et al. (2019) verified Freesurfer's cortical segmentation; Abdallah et al. (2016) assessed the manual segmentations of diffused low-grade gliomas; Valindria et al. (2017), Robinson et al. (2017), and Robinson et al. (2019) proposed QC schemes on cardiovascular datasets. Our approach is the first one to deal with QC of MRI CC segmentations, assigning, for each segmentation, a probability (quality score) of belonging to one of two classes: correct or incorrect. By applying a decision threshold to this probability, it is possible to label a new segmentation as correct or incorrect.

Selection of the proper decision threshold is a trade-off that depends on the specific application. In practice, lower decision threshold values are preferable because they increase the detection capability (increase TPR) at the cost of increasing false positives. However, as incorrect segmentations, including false positives, go on to a posterior manual verification/correction after QC is applied, it is possible to guarantee the effectiveness of the QC stage. Fortunately, in our framework, TPR and TNR remain high for a wide range of decision threshold values, thanks to the high AUC obtained (Fig. 7b).

We defined the decision threshold using the $F1_{score}$ resulting in 2 false positives and 7 false negatives (Fig. 8). The false positive cases could be avoided by using a more traditional decision threshold, such as 50%. On the other hand, the false negatives presented low output probabilities. Five out of these 7 cases resulted from masks that are divided into two sections (Fig. 8b), and the shape signature, when evaluated in one of them, seemed normal. These cases could be avoided by adding a preliminary verification step that rejects two-portion segmentations, using a connected-component algorithm for example.

The shape signature was a robust and versatile descriptor for CC segmentation quality assessment. As the signature is extracted at various resolutions, it allows an assessment of the CC segmentation at different levels (fine detail and coarse shape). However, it is challenging to choose the proper combination of resolutions to accomplish the proposed classification task. By constructing an ensemble of classifiers according to high-quality and diversity criteria, we achieved AUC = 98.25%, a notable score for classification tasks. Rejecting

Table 3
New values tested, one by one, for the hyper-parameters of the ensemble. Each row is an experiment where is changed one hyper-parameter value. For comparison, the AUC base, with the original values, obtained 98.25% (last row).

R	Quality threshold	Dissimilarity Measure	Criterium representative	Final AUC
25	0.90	Disagreement	Intra-cluster distance	97.16%
50	**0.80**	Disagreement	Intra-cluster distance	96.87%
50	**0.93**	Disagreement	Intra-cluster distance	93.57%
50	0.90	**Q statistics** (Kuncheva and Whitaker, 2003)	Intra-cluster distance	97.31%
50	0.90	Disagreement	**Best AUC**	98.22%
50	0.90	Disagreement	**Random**	97.32%
50	0.90	Disagreement	Intra-cluster distance	98.25%

Bold value signifies specific parameter that was changed in every experiment.

individual classifiers with low individual AUC was essential to exclude low-quality components of the final ensemble (Fig. 6a). The optimal ensemble was constructed, allowing the automated selection and agreement of 12 individual classifiers spread along the resolutions.

The optimal ensemble led to the best final AUC, keeping the number of classifiers low (twelve). The experiments showed poor performance with fewer components (below 12 components), and there was no substantial improvement by using more components (above 12 components) (Fig. 7a). The optimal ensemble also obtained a higher AUC (98.25%) than the ensemble with the 12 best-quality individual classifiers (96.81%), demonstrating that diversity is important in building the ensemble (Fig. 9).

Experiments (Section 3.4) confirmed the robustness of our framework to the selection of hyper-parameters (Table 3). The most relevant hyper-parameter was the quality threshold. When this value was either increased or decreased, the final result got worse, showing that a quality threshold value of 0.90 is an optimal choice.

The use of the framework for the analysis of other T1-MRI CC segmentation datasets is straightforward since we made the Python script and the trained classification model available. Given that the framework was trained in two large datasets and it learned the characteristic shape of the CC, it does not need ground-truth for quality assessment of new segmentations. Also, it can be used in other MRI sequences such as T2-MR images and Diffusion-Weighted images, where the CC preserves its characteristic shape. In any of these cases, the hyper-parameter values can be kept the same as the original ensemble (the Python script and the trained model are by default configured with these values Table 3). The only value that needs to be defined is the final decision threshold for classifying input segmentations as correct or incorrect.

There are specific populations such as fetal, newborn (Huang et al., 2006), or non-disabled elderly populations (Ryberg et al., 2007) in which the CC shape changes in relation to the dataset used in this work. In these cases, the use of our framework probably requires re-training the supervised model in the target image dataset and setting the hyper-parameters to achieve good QC results. On the other hand, it is expected that our framework will not perform properly, even if re-trained, in the presence of CC malformations, tumors, and agenesis (Hetts et al., 2006). In these cases, it is difficult for the classifier to learn the separation model between correct and incorrect classes.

Finally, our method can be extended to any segmentation QC application where the masks to be evaluated have a generic shape along with the dataset. This is the case for several medical imaging applications, such as brain MRI, in which structures and organs maintain their shape along with the population. In these cases, it is possible to customize the shape signature to grasp the generic shape along the target dataset. The ability of our framework to adapt to several applications occurs because the multi-resolution descriptor and the supervised ensemble ensure the framework learns the best characteristics of the shape to perform the proposed classification task.

5. Conclusions

In large-scale imaging studies, where segmentation is a mandatory step, the automatic detection of incorrect segmentations is of utmost importance. In this work, we proposed a framework for QC of CC segmentations that does not need ground-truth.

Our framework used a multi-resolution shape descriptor and automatic selection of individual classifiers through a clustering technique. Since the selection of the proper resolutions for QC is tricky, the use of an ensemble yielded a better solution. Two criteria led to the construction of the optimal ensemble: selection of high-quality classifiers and variability between them. This scheme was used to build an ensemble of 12 components with various resolutions, achieving an AUC of 98.25% on the test set containing 207 segmentations.

Finally, since our framework is based on a shape descriptor and does not depend on intensities, its use can be easily extended to assess segmentations done over other MRI sequences, such as T2-MR images and Diffusion-Weighted images. The method can be extended to other QC segmentation applications by training the framework in the target dataset.

Author contributions

William Javier García Herrera: Conceptualization, Methodology, Software, Investigation, Data Curation, Writing – Original Draft, Visualization.

Mariana Eugênia de Carvalho Pereira: Validation, Data Curation, Writing – Review Editing.

Mariana Pinheiro Bento: Validation, Data Curation, Writing – Review Editing.

Aline Tamires Lapa: Resources, Data Curation.

Simone Appenzeller: Resources, Data Curation, Funding acquisition.

Letícia Rittner: Conceptualization, Methodology, Software, Validation, Resources, Data Curation, Writing – Original Draft, Supervision, Funding acquisition.

Conflict of interest statement

We have no conflicts of interest to declare.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP – process CEPID 2013/07559-3) and by the National Counsel of Technological and Scientific Development (processes 190557/2014-1 and 308311/2016-7). We would also like to thank Prof. Heath Pardoe, Associate Professor at the Department of Neurology of the NYU Langone Medical Center, New York University, for providing us with the automatic and manually corrected segmentations from the ABIDE database.

References

Abdallah, M.B., Blonski, M., Wantz-Mézières, S., Gaudeau, Y., Taillandier, L., Moureaux, J.-M., 2016. Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset. IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC) 4403–4406.

Adamek, T., O'Connor, N.E., 2004. A multiscale representation method for nonrigid shapes with a single closed contour. IEEE Trans. Circuits Syst. Video Technol. 14, 742–753.

Ardekani, B., 2013. Yuki Module of the Automatic Registration Toolbox (Art) For Corpus Callosum Segmentation. http://www.nitrc.org/projects/art.

Ardekani, B.A., Kershaw, J., Braun, M., Kanuo, I., 1997. Automatic detection of the mid-sagittal plane in 3-d brain images. IEEE Trans. Med. Imaging 16, 947–952.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2, 121–167.

Chao, L.-C., Chien, C.-F., 2010. A model for updating project s-curve by using neural networks and matching progress. Autom. Constr. 19, 84–91.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Cover, G., Herrera, W., Bento, M., Appenzeller, S., Rittner, L., 2018. Computational methods for corpus callosum segmentation on MRI: a systematic literature review. Comput. Methods Program. Biomed. 154, 25–35.

Faria, F.A., Santos, J.A., Rocha, A., Torres, R.S., 2012. Automatic classifier fusion for produce recognition. Proc. of XXV SIBGRAPI – Conf. on Graph., Patterns and Images.

Gordillo, N., Montseny, E., Sobrevilla, P., 2013. State of the art survey on MRI brain tumor segmentation. Magn. Reson. Imaging 31, 1426–1438.

Hajian-Tilaki, K., 2013. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. Casp. J. Intern. Med. 4, 627.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143, 29–36.

He, Q., Duan, Y., Miles, J., Takahashi, N., 2007. A context-sensitive active contour for 2D corpus callosum segmentation. Int. J. Biomed. Imaging 2007.

Herrera, W., Bento, M., Rittner, L., 2019. Corpus Callosum Shape Signature for Segmentation Evaluation, XXVI Brazilian Congress on Biomedical Engineering. IFMBE Proceedings 70/2, 143–147.

Hetts, S.W., Sherr, E.H., Chao, S., Gobuty, S., Barkovich, A.J., 2006. Anomalies of the corpus callosum: an mr analysis of the phenotypic spectrum of associated malformations. Am. J. Roentgenol. 187, 1343–1348.

Hiess, R.K., Alter, R., Sojoudi, S., Ardekani, B., Kuzniecky, R., Pardoe, H., 2015. Corpus

callosum area and brain volume in autism spectrum disorder: quantitative analysis of structural mri from the abide database. J. Autism Dev. Disord. 45, 3107–3114.
Hofer, S., Frahm, J., 2006. Topography of the human corpus callosum revisited-comprehensive fiber tractography using diffusion tensor magnetic resonance imaging. NeuroImage 32, 989–994.
Huang, H., Zhang, J., Wakana, S., Zhang, W., Ren, T., Richards, L.J., Yarowsky, P., Donohue, P., Graham, E., van Zijl, P.C., et al., 2006. White and gray matter development in human fetal, newborn and pediatric brains. Neuroimage 33, 27–38.
Huang, C., Wu, Q., Meng, F., 2016. Qualitynet: segmentation quality evaluation with deep convolutional networks. Visual Communications and Image Processing (VCIP) 1–4.
Jomma, H.D., Hussein, A.I., 2016. Circle views signature: a novel shape representation for shape recognition and retrieval. Can. J. Electr. Comput. Eng. 39, 274–282.
Klapwijk, E.T., Van De Kamp, F., Van Der Meulen, M., Peters, S., Wierenga, L.M., 2019. Qoala-t: a supervised-learning tool for quality control of freesurfer segmented mri data. NeuroImage 189, 116–129.
Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: a review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. 160, 3–24.
Kuncheva, L.I., Whitaker, C.J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. 51, 181–207.
Mokhtarian, F., Mackworth, A.K., 1992. A theory of multiscale, curvature-based shape representation for planar curves. IEEE Trans. Pattern Anal. Mach. Intell. 14, 789–805.
Oliphant, T.E., 2006. A guide to NumPy, vol. 1 Trelgol Publishing, USA.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Peng, B., Zhang, L., Mou, X., Yang, M.-H., 2017. Evaluation of segmentation quality via adaptive composition of reference segmentations. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1929–1941.
Powers, D.M., 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation.
Reeves, A.P., Liu, S., Xie, Y., 2016. Image segmentation evaluation for very-large datasets. Proc. SPIE 9785 9785-11.
Robinson, R., Valindria, V.V., Bai, W., Suzuki, H., Matthews, P.M., Page, C., Rueckert, D., Glocker, B., 2017. Automatic quality control of cardiac mri segmentation in large-scale population imaging. International Conference on Medical Image Computing and Computer-Assisted Intervention 720–727.
Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., et al., 2019. Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study. J. Cardiovasc. Magn. Reson. 21, 18.
Ryberg, C., Rostrup, E., Stegmann, M.B., Barkhof, F., Scheltens, P., van Straaten, E.C., Fazekas, F., Schmidt, R., Ferro, J., Baezner, H., et al., 2007. Clinical significance of corpus callosum atrophy in a mixed elderly population. Neurobiol. Aging 28, 955–963.
Shi, R., Ngan, K.N., Li, S., Paramesran, R., Li, H., 2015. Visual quality evaluation of image object segmentation: subjective assessment and objective measure. IEEE Trans. Image Process. 24, 5033–5045.
Shi, W., Meng, F., Wu, Q., 2017. Segmentation quality evaluation based on multi-scale convolutional neural networks. In: Visual Communications and Image Processing (VCIP). IEEE. IEEE, pp. 1–4.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437.
Tulyakov, S., Jaeger, S., Govindaraju, V., Doermann, D., 2008. Review of classifier combination methods. Mach. Learn. Document Anal. Recognit. 361–386.
Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2017. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans. Med. Imaging 36, 1597–1606.

5 Quality control of corpus callosum segmentations: deep learning approach

In the previous chapter, we presented the first QC framework applied to CC segmentations using manually extracted features (shape signature) and classical machine learning techniques, with outstanding performance ($AUC = 98.25\%$) on T_1-MRI. At the end of the paper, we hypothesized about two new scenarios: diffusion MRI and tumor patients, where this framework would have both advantages and disadvantages, given its exclusive dependence on the segmentation mask shape.

In this chapter, we present our work entitled **Automatic quality control on corpus callosum segmentation: Comparing deep and classical machine learning approaches** which was submitted to the Neurocomputing journal. It makes two significant contributions: proposal of a QC method for CC segmentations based on deep learning and comparison of this new method with the classical machine learning framework in the cases listed in the previous paragraph.

The CNN method achieved a similar performance ($AUC = 97.98\%$) to that of the classical machine learning approach on the test dataset. When comparing both methods, the classical approach got the best result in diffusion MRI while the CNN performed better in the tumor dataset. The main advantage of the deep learning approach is its capability to extract features of the mask and the image, gaining knowledge about the context of the segmentation. However, as this approach is sensitive to the image intensity, it is not easy to generalize it to domains with different intensity (e.g., T_1 to diffusion). In this case, a domain adaptation technique could be useful to improve the CNN performance on this new domain.

Some additional techniques were tested along the experiments including: random variation of brightness and contrast for data augmentation, blurring the segmentation mask using a Gaussian function, and usage of the CNN ResNet50. However, none of these techniques improved the performance of the network substantially and therefore they were not included in the final model.

Automatic quality control on corpus callosum segmentation: Comparing deep and classical machine learning approaches

William Garcia Herrera[§], Simone Appenzeller[†], Fabiano Reis[‡], Danilo Rodrigues Pereira[*], Leticia Rittner[§]

[§] *Medical Image Computing Laboratory (MICLab), School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Brazil*
[†] *Rheumatology Unit, Department of Medicine, School of Medical Science, University of Campinas (UNICAMP), Brazil*
[‡] *Radiology Department, School of Medical Science, University of Campinas (UNICAMP), Brazil*
[*] *Pathophysiology Graduate Program, School of Medical Science, University of Campinas (UNICAMP), Brazil*

Abstract

The Corpus callosum (CC) is a massive white matter structure in the brain, and changes in its shape and volume are associated with subject characteristics, several diseases, and clinical conditions. The CC is mostly studied in magnetic resonance imaging (MRI), where it is segmented to extract valuable information. With the increasing availability of MRI data and the proliferation of automated algorithms to perform CC segmentation, quality control (QC) verification is mandatory to assure reliability in the entire analysis pipeline.

We propose a convolutional neural network (CNN) for QC of CC segmentations. The CNN gets information on the mask and contextual information on the image and performs deep feature extraction using a pre-trained model. The CNN model was fine-tuned using T_1-MRI images with CC masks, in the task of classifying correct or incorrect segmentations.

The CNN-based approach got an area under the curve (AUC) of 97.98% on the test set. To validate our proposal, we compared it with a classical machine learning approach, based on a SVM ensemble, trained in the same task. The classical approach got the best performance in the diffusion domain, but the CNN overcame it in subject with tumor affecting the CC. Both approaches were compared in diffusion MR images and patients with tumor to test generalization capability to other domains.

The simple CNN architecture got similar performance as the classical machine learning approach, in the dataset used to train the models. However, the CNN resulted more versatile than the classical machine learning model, better able to adapt to unseen patients with tumor, and best suited to learn other patterns with domain adaptation.

Keywords: corpus callosum, segmentation, quality control, convolutional neural network, magnetic resonance imaging

1. Introduction

The Corpus callosum (CC) is the largest white matter structure in the brain and is responsible for the inter-communication of the brain hemispheres [1]. The CC is essential in medical, clinical and, research areas since changes in its shape and volume are associated with subject

[†] Corresponding author. Medical Image Computing Laboratory (http://miclab.fee.unicamp.br/). E-mail address: w162642@dac.unicamp.br (W. G. Herrera).

characteristics, several diseases, and clinical conditions [2]. The CC is mostly studied in magnetic resonance imaging (MRI), in which its segmentation is required [3]. Segmentation is the process of dividing an image into non-intersecting and homogeneous regions. This process is the preliminary step to several analysis stages and can be accomplished by using an automatic, semi-automatic, or manual segmentation method [4].

With the increasing availability of MRI data, manual or even semi-automatic methods are unfeasible, with automatic methods being the only way to perform segmentation on a large scale [5]. Because automatic methods are not fully reliable, quality control (QC) verification is mandatory to assure reliability in the entire analysis pipeline. QC stage allows detecting incorrect segmentation in order to be discarded or revised, avoiding the introduction of errors into the remaining pipeline [6].

In medical imaging, efforts have been concentrated on proposing and improving automated segmentation methods, and so there are few works in QC of segmentation. Among QC proposals, Bouix et al. [7] presented three techniques - common index agreement, expectation-maximization, and multidimensional scaling - for evaluating brain tissue segmentation. Abdallah et al. [8] used a statistical model to assess the manual segmentation of diffused low-grade gliomas on MRI by comparing extracted measurements from the segmentation with a reference. A random forest model developed by Klapwijk et al. [9] evaluated cortical Freesurfer segmentation on MRI.

Valindria et al. [10] and Robinson et al. [11] used a deep learning approach based on reverse classification accuracy (RCA) to evaluate segmentation accuracy in the lack of ground-truth. This work was then tested on a large MRI cardiovascular dataset [6]. Roy et al. [12] proposed a Bayesian extension of the QuickNat CNN for delivering both the segmentation with the voxel-wise uncertainty map and the final quality control measure. This approach was trained and tested in four small datasets for different brain structures on MRI.

In our previous work [13], we presented a framework for QC of CC segmentation on MRI using an ensemble of support vector machine (SVM). It is based on a shape descriptor, depending only on the shape of the segmentation mask, making it independent of which MRI sequence the segmentation was performed on. While it is an advantage because the same framework can be applied in other MRI sequences, it may be seen as a shortcoming, since it operates decoupled from the MR image, missing the contextual information of the image. In the latter situation, subjects with tumor, correct but shifted masks (e.g., registration errors) and specific populations (e.g., fetal, newborn, or elderly people) may be misclassified.

On medical imaging, deep learning applications have achieved remarkable outcomes in various applications: classification of exams, illness and lesions; detection of tumors, organs, regions and landmarks; segmentation of organs and lesions; registration; big data applications such as content-based image retrieval and combination with reports; and generative models for enhancing, de-noising, normalizing and pattern discovery [14, 15]. Deep learning is a group of machine learning techniques, in which many layers of information processing are stacked for performing complex pattern recognition and feature representation tasks. Deep learning allows constructing complex concepts out of simpler ones by extracting information at increasing levels of detail through simple operations, such as convolution [16].

Among the most popular deep learning techniques, convolutional neural networks (CNN) are the most used to deal with images, overshadowing classical machine learning methods on several applications. Therefore they are a promising approach when dealing with QC of image segmentation. However, CNN are not the solution for every case and require more data and computational resources, leaving room for classical machine learning approaches [17]. Therefore, we propose a method for QC of CC segmentations using a CNN. We compare its performance with the performance of the classical machine learning method, and finally, we extend the trained models to other domains: diffusion MRI sequence and patients with tumor. Testing in other datasets allows us to verify our model's generalization capability and possibly allows

its use in other applications. Therefore, this paper's contribution is twofold: to propose a QC method for CC segmentations using a CNN and to verify its performance in real case scenarios compared to the classical machine learning framework.

This paper is divided into four sections, as follows: section 2 introduces deep learning and CNN, and explains the proposed model; section 3 presents the performance of the CNN on the test dataset, and the comparison results with a classical machine learning approach; section 4 examines the results in terms of applicability, advantages, and limitations of our proposal; and section 5 summarizes and remarks the main points of our work.

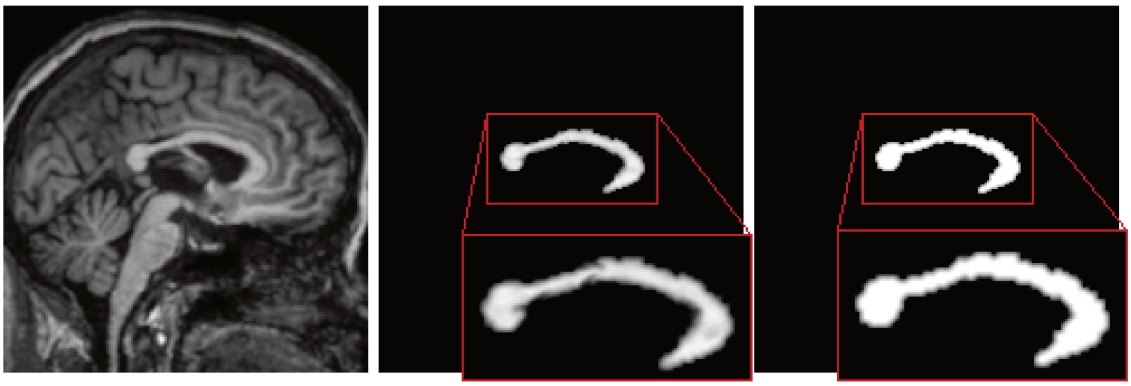
2. Methods

In this work, the QC problem is approached by classifying the segmentation as correct or incorrect. The probability of the segmentation being incorrect is the quality control score (QCS), ranging from 0% (completely correct) to 100% (completely incorrect). For this purpose, we used one of the first proposed CNN: the residual learning network (ResNet) [18]. In order to make the model more effective, we used some techniques that will be explained in sequence: section 2.1 describes how the input was arranged to evaluate a binary mask in conjunction with the original image; section 2.2 expands the ResNet architecture used; and section 2.3 explains the process carried out to perform the network training.

2.1. Input arrange

When evaluating quality in segmentation, the binary segmentation mask must be considered. In the classical machine learning approach, previously proposed, features are extracted manually from the segmentation, missing all the contextual information from the original image. In contrast, this CNN-based approach includes the image and its segmentation, using information of the available context.

The image and the binary mask were organized into a three-channel arrangement: MR intensity image (Fig. 1a), MR image multiplied by the binary mask (Fig. 1b), and the mask itself (Fig. 1c). This arrangement allows taking advantage of the image and the mask together, and the central channel highlights the contextual information extracted by the mask.



(a) (b) (c)

Figure 1: Three-channel arrangement for the image and the binary mask: (a) image, (b) image multiplied by the binary mask, (c) binary mask

2.2. CNN architecture

CNN networks make possible feature extraction based on deep-fashion arrangements. Resnet is a popular CNN architecture that reformulated the layers as learning residual functions. These

residual representations are easier to optimize and allow gain accuracy on behalf of deeper architectures. In practice, the residual representations are achieved by inserting shortcut connections every two convolutional layers [18].

We used ResNet18, which has one initial block, comprising one convolutional layer, and eight sequential blocks, composed of two convolutional layers. A batch normalization layer follows all convolutional layers, and Relu activations are used inside the sequential blocks, between the convolutional layers. Right after the convolutional blocks, three fully connected layers were inserted, providing the QCS.

2.3. Training process

Since the used CNN is a supervised model, it requires a training process to learn the distribution of samples, discovering the classification hyperplane that best separates the classes. Deep learning models are recognized for requiring both computing power and large datasets during the training phase. To alleviate this, we used transfer learning to take advantage of pre-trained CNN, achieving an effective feature extractor [19]. In this paper, we used ResNet18, pre-trained on ImageNet [20]. In order to adapt the ImageNet RGB domain to our problem, we froze the convolutional part, except for the last block, and adjusted, by fine-tuning, this final block and the final fully connected layers [21]. We also used the data augmentation technique, which increases generalization performance, focusing on the training dataset. The idea is to extract more information from the original dataset through random augmentations [22]. In our case, we applied random affine rotations. Since the rotation left some pixels in the border with unknown value, we made a crop in the center of the image, discarding the border. Because the CC is always in the center of the brain, this crop did not affect it.

3. Results

We used 907 MR images from two distinct acquisition sequences, and the CC was segmented using different methods (Table 1). All data used in this work was fully anonymized and approved by the local ethical committee. All of the participants were adequately informed, and they agreed to participate in the studies.

Experiment	MRI sequence	Study	Samples	Segmentation method
CNN Evaluation / Performance comparison	T_1-MRI	Unicamp	397	Freesurfer [23]
		Unicamp	151	Manual
		ABIDE [24]	247	Yuki [25]
Generalization (FA)	Diffusion-MRI	Unicamp	105	ROQS [26]
Generalization (Tumor)	T_1-MRI	Tumor	6*	Manual
		Tumor	4*	Freesurfer

*3 subjects were segmented with both methods, Manual and Freesurfer.

Table 1: MRI datasets used in CNN experiments: normal subjects T_1-MRI, normal subjects Diffusion-MRI, and T_1-MRI of subjects with tumor.

The experiments were performed using Python [27] along with Pytorch [28] for the CNN implementations. For reproducibility purposes, the trained model and the source code are available on GitHub[1].

[1] https://github.com/wilomaku/CC_QC_CNN

In the first experiment, we tested the CNN-based method using a T_1-MRI dataset (Section 3.1), where three different methods segmented the CC. In the second experiment, its performance was compared to the performance of our previous QC framework based on classical machine learning (Section 3.2). Finally, we verify the generalization of both models to other domains: diffusion MRI and tumor datasets (Section 3.3).

3.1. Evaluation of the CNN-based proposed method

For training and testing the proposed model, we performed the experiments over the T_1-MRI dataset exclusively. From each T_1-MRI volume, the mid-sagittal slice was extracted by using the *acpcdetect* tool from the ART toolbox [29]. Then, some samples of the dataset were discarded due to one of the following reasons: significant different image size, wrong selection of the mid-sagittal slice, or unclear accuracy of segmentation because of very subtle errors. The final dataset comprised 685 samples distributed into 400 correct segmentations (58%) and 285 incorrect segmentations (42%). Furthermore, the dataset was stratifiedly partitioned (Table 2).

Samples	Proportion	Use
439	64%	Training
110	16%	Validation
136	20%	Testing

Table 2: T_1-MRI dataset distribution used in CNN experiments comprising training, validation and testing sets.

Each sample of the dataset is composed of a 2D image and its associated segmentation mask. Because the image size varies with the studies, we established a 240x240 size to ensure that the network worked with any input. Images with different sizes were re-scaled using the Lanczos interpolation, a high-quality convolutions-based algorithm with the Lanczos kernel [30]. Since the network was pre-trained on ImageNet, input should be normalized with per-channel mean ($[0.485, 0.456, 0.406]$) and standard deviation ,($[0.229, 0.224, 0.225]$) values. Data augmentation was applied in the form of random rotations between $-3°$ to $3°$. Afterward, a central 176x176 crop was performed.

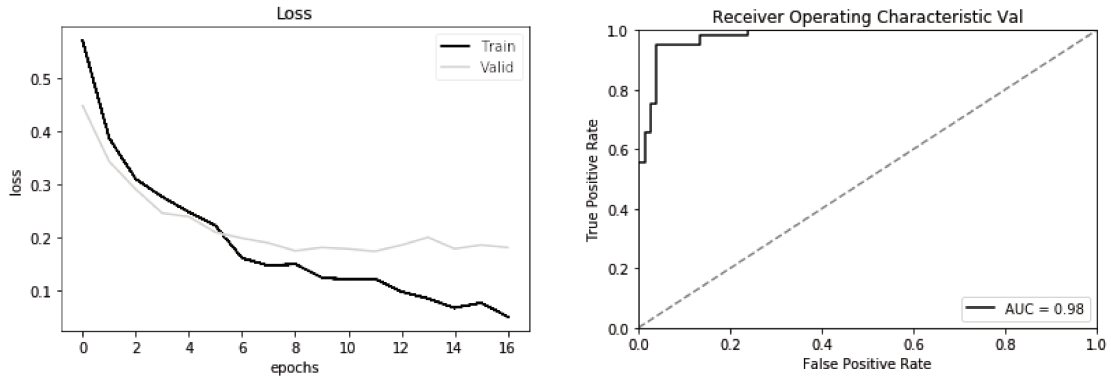
The ResNet18 was configured as specified in section 2.2. Three fully connected layers that generated the QCS were added at the end of the network, with sizes 1024, 512, 2, respectively.

We trained the model along 50 epochs with cross entropy loss function, Adam optimizer [31], a learning rate of $1x10^{-5}$, and a mini-batch size of 16. After 12 epochs, the network suffered overfitting, and after 5 additional epochs (17 epochs) with no improvement, the training was halted using early-stop. The state of the network (learned parameters) was saved at the 12 epoch, achieving 0.121 and 0.173 of cross entropy loss (Fig. 2a).

The AUC for the testing set was 97.98% (Fig. 2b). The decision threshold, defined at the maximum $F1_{score}$, was 53% to classify a given segmentation as correct or incorrect. With this decision threshold, the CNN had seven miss-classifications (accuracy=94.85%) (Fig. 3).

3.2. Performance comparison of QC methods

The final AUC of our previous classical approach [13] was similar to our current work (Section 3.1). Therefore, in this section, we propose a performance comparison in terms of training and execution time and dataset size. These two aspects have a direct impact on the daily use and the final AUC of the method, respectively, and can determine the choice of one of them over the other. We compared the CNN-based approach with our previous one, using an SVM ensemble [13], in two experiments: time measurement for both training and testing stages (Section 3.2.1) and impact of training set size on the final result (Section 3.2.2).



(a) (b)

Figure 2: Performance of the CNN-based QC method : (a) Training and validation loss for ResNet18 along 17 epochs, and (b) ROC and AUC for the test set.

(a) (b) (c)

Figure 3: Confusion matrix and false result cases for the CNN at *threshold* = 53%: (a) confusion matrix, (b) false negative example: incorrect segmentation (true label = 1) with $p = 7.9\%$, (c) false positive example: correct segmentation (true label = 0) with $p = 96.8\%$. The segmentation mask is depicted as the shaded red region.

3.2.1. Time comparison

We measured time using the same previous sets (Table 2) at training and testing stages for both models: SVM ensemble and CNN-based (Table 3).

It must be remarked that the training stage was not performed in the same hardware as the test stage. Furthermore, a GPU was used for training the CNN to accelerate the process. On the other hand, the hardware was the same at the test stage for both SVM ensemble and CNN models.

3.2.2. Variation of training dataset size

To verify the impact of the training dataset size on the final performance (Fig. 4), we held the testing dataset fixed at 20% of all samples and trained both models with different amounts of training samples from the T_1-MRI dataset (Table 2). For every chosen dataset size, we used the same proportion of 4 training samples to 1 validation sample, and the experiment was executed 5 times, with their mean value as final registered QCS.

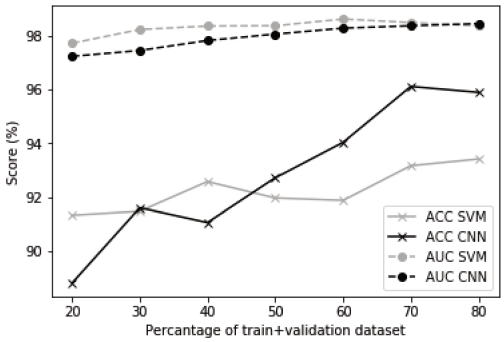
3.3. Generalization to other domains

Supervised machine learning models are data-driven and suffer from a lack of generalization, presenting a drop in performance on new unseen datasets. This is particularly common in medical imaging, when the subjects of the testing dataset present some pathology never seen in the training samples. To test the model's generalization to other domains, we used the

Stage	CPU Model	# Cores	Memory	Model	GPU	Time (mins)
Training	Intel Xeon (2.30GHz)	1 (2*)	11.98 GB	SVM ensemble	NA	28'06"
				CNN-based	12 GB**	140'36"
Testing	Intel i7-8750H (2.20GHz)	6 (2*)	14.68 GB	SVM ensemble	NA	0'9.75"
				CNN-based	NA	0'9.48"

*Threads per core.
**Nvidia Tesla K80.

Table 3: Time comparison among training and testing stages for the SVM ensemble (classical machine learning) and the CNN (deep learning) models.



(a)

%	# Training / Validation samples	AUC (%) SVM	AUC (%) CNN	ACC (%) SVM	ACC (%) CNN
20%	110/27	97.72	97.23	91.32	88.80
30%	165/41	98.23	97.45	91.47	91.60
40%	219/55	98.36	97.82	92.57	91.05
50%	274/68	98.37	98.06	91.96	92.71
60%	329/82	98.62	98.28	91.87	94.03
70%	383/96	98.49	98.37	93.16	96.11
80%	438/110	98.37	98.45	93.41	95.89

(b)

Figure 4: Comparison of AUC and accuracy (ACC) for different number of training/validation dataset samples and both models: SVM ensemble and CNN. The test dataset was fixed to 136 samples (20% of the total dataset) for all the experiments.

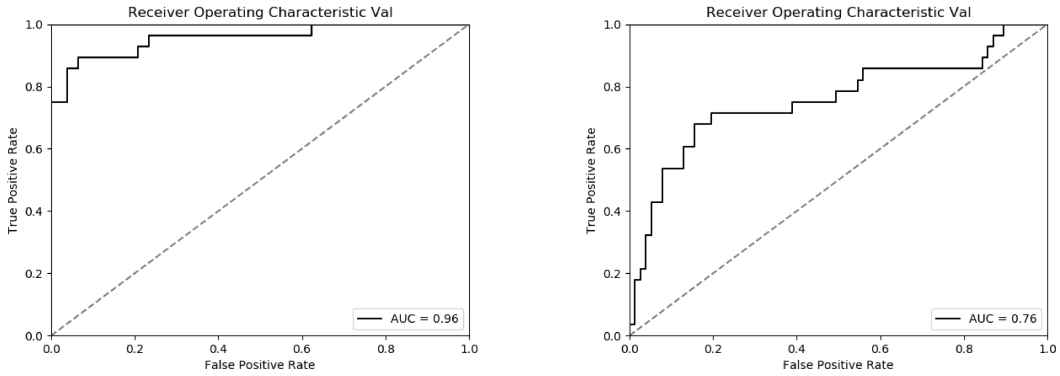
trained models in the T_1-MRI dataset, with no adaptation or fine-tuning, and tested them on two datasets: diffusion-MRI (Section 3.3.1) and patients with tumor (Section 3.3.2).

Since the SVM model requires only the segmentation masks and is invariant to the image size, the input masks of the two testing datasets were used directly, requiring no changes. On the other hand, the CNN-based approach requires a fixed size input; therefore, the images were cropped at their center, remaining at 232x232, and then re-scaled to 240x240 using Lanczos interpolation. For these experiments, we fixed the separation threshold of the QCS in 50%, which means scores above 50% were considered incorrect (True), and scores below were considered correct segmentations (False).

3.3.1. Diffusion MRI (FA images)

Although T_1 is the most used MRI sequence for studying the CC, its segmentation on Diffusion-MRI and its maps, such as fractional anisotropy (FA), is important, giving information about the microstructure and fibers characteristics based on measurements of water molecules' movement along the tissues [32]. Also, automated and semi-automated methods for CC segmentation on diffusion-MRI exist [33–38], despite its low resolution and complexity. Therefore, checking the quality of the segmentations on diffusion-MRI is essential.

To test both models on diffusion-MRI, we used a dataset composed of 105 images of FA, segmented using the ROQS algorithm (Table 1). The SVM-based approach had better AUC (95.55%) and accuracy (88.57%) than the current proposal (AUC = 75.70% / $accuracy$ = 78.10%) (Figure 5 and Table 4).



(a) (b)

Figure 5: ROC and AUC for the diffusion MRI dataset: (a) SVM ensemble, and (b) CNN-based approach.

Approach	True positive	True negative	False positive	False negative	Accuracy
SVM ensemble	25	68	9	3	88.57%
CNN	9	73	4	19	75.70%

Table 4: Confusion matrix results for the diffusion MRI dataset.

3.3.2. Patients with tumor

In this experiment, we used the trained models and tested them with 7 subjects that had tumor deforming the CC, either directly (tumor is on the CC) or indirectly (tumor is around it, displacing nearby tissues and affecting its shape and/or position). The tumor cases evaluated are malignant gliomas, i.e., that arose from the glial tissue, with a high growth index. Three out of seven patients had a mid-grade tumor (grade III), known as anaplastic astrocytoma, while the remaining four had high-grade tumors (grade IV), the highest grade gliomas, called glioblastomas [39]. Since the patient's head symmetry was affected by the tumor, the mid-sagittal slice selection was made manually. The CC in T_1-MR images was segmented using Freesurfer (1 subject), manual segmentation (3 subjects), and both methods (3 subjects), resulting in 10 segmentations (Table 5 and Fig. 6).

Overall, the classical approach had $AUC = 70.83\%$ / $accuracy = 60.0\%$), while our current proposal had ($AUC = 83.33\%$ / $accuracy = 80.0\%$).

4. Discussion

In this work, we proposed the use of a CNN for performing QC on CC segmentations. The arrangement of the input channels allows obtaining information of the segmentation itself (first channel), contextual knowledge of the CC region where the segmentation mask is inserted (second channel), and knowledge about the mask segmentation (third channel). The network's architecture remained simple, with ResNet18 becoming one of the first deep learning approaches reporting good performance on classification tasks. This made it possible to obtain a model with similar performance to our previous classical machine learning approach ($AUC = 98\%$). It should be noted that the test set was not the same since we wanted to train the CNN with a larger part of the dataset. It went from 30%, in the classical approach, to 20%, in our current work. Nevertheless, in both cases, it was randomly sampled, and therefore the final AUC is comparable. Furthermore, as noticed in the variation of the training dataset size experiment (Figure 4), the impact of this variation on the final AUC was small for both

Subject	Method	Ground-truth	SVM ensemble		CNN	
			QCS	Predict	QCS	Predict
1	Manual	Correct	99.61	**Incorrect**	91.48	**Incorrect**
	Freesurfer	Incorrect	99.60	Incorrect	91.50	Incorrect
2	Manual	Correct	95.75	**Incorrect**	1.73	Correct
	Freesurfer	Incorrect	99.59	Incorrect	72.12	Incorrect
3	Manual	Correct	22.13	Correct	0.6	Correct
	Freesurfer	Incorrect	27.18	**Correct**	5.36	**Correct**
4	Manual	Correct	45.78	Correct	1.15	Correct
5	Manual	Correct	78.07	**Incorrect**	0.67	Correct
6	Manual	Correct	17.39	Correct	23.37	Correct
7	Freesurfer	Incorrect	99.61	Incorrect	57.95	Incorrect

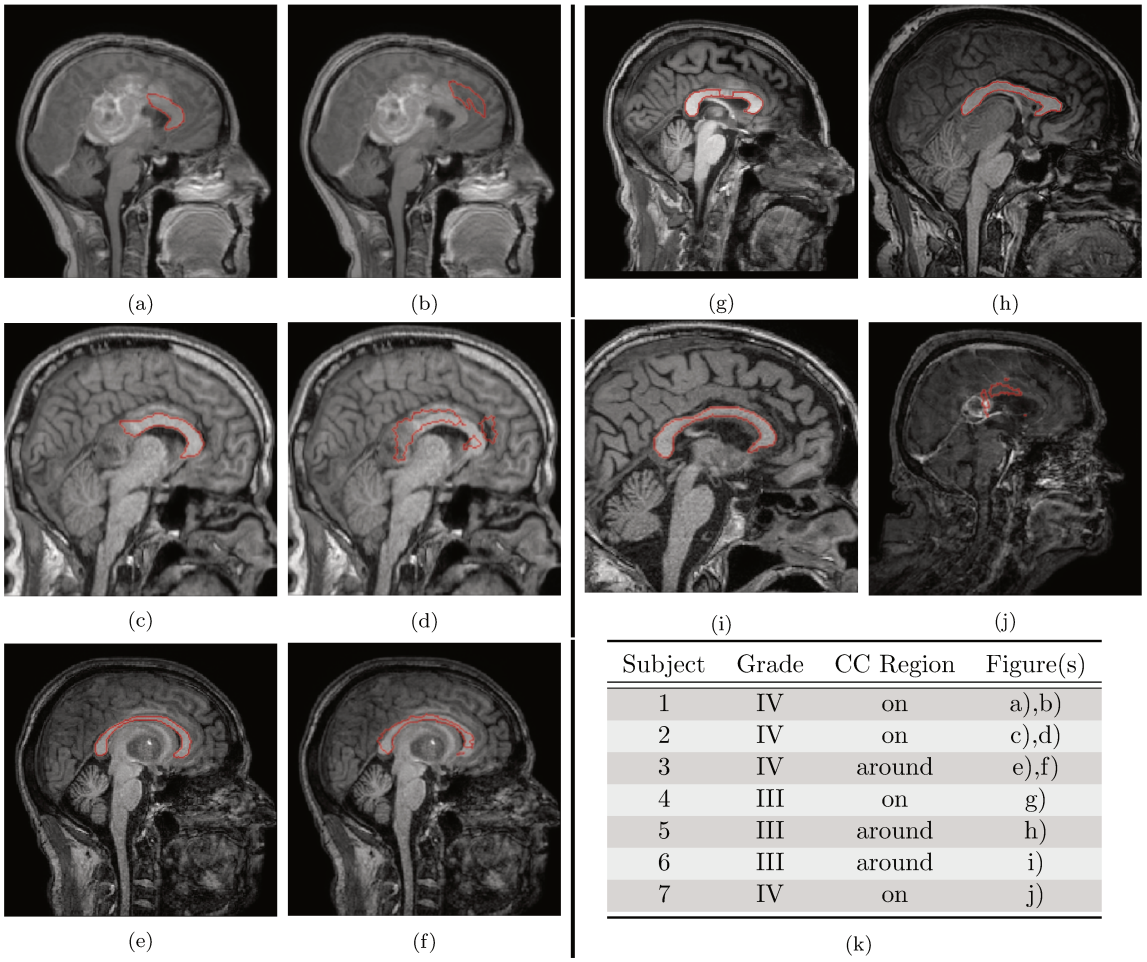
Table 5: QCS for the SVM ensemble and CNN approaches for patients with tumor affecting the CC. Two methods were evaluated: Manual and Freesurfer. The predict outcome corresponds to a threshold separation of 50%.

approaches. This might seem surprising, especially in CNN, which are considered eager for large amounts of information. However, the implementation of transfer learning and fine-tuning strategies alleviated this requirement. Nevertheless, increasing the size of the training dataset did help, improving the accuracy in the CNN-based approach, and allowing outperform the classical machine learning approach.

By selecting a simple pre-trained model such as the ResNet18, in order to construct the quality control system, we obtained fair training and testing times using moderate hardware (Table 3). Training the CNN in a non-last generation GPU (Nvidia Tesla K80, launched in 2014) took almost 5 times longer than the classical approach. The pre-training strategy mitigated the training time because the CNN converged in just 12 epochs. More importantly, due to the practical use of our quality control system, the testing time to evaluate 136 subjects was less than 10 seconds, that is, the same as the classical approach and fast enough to be used in any real scenario with large datasets.

To compare both approaches in new domains, we tested the generalization of the models, trained on standard T_1-MR images, to new real scenarios: diffusion FA images and tumor cases. In the diffusion domain, we tested over 105 new subjects segmented with a semi-automatic method. The SVM-based approach performed better ($AUC = 95.5\%$) than the CNN-based approach ($AUC = 75.7\%$). When switching the domain, from T_1 to diffusion MRI, the CC shape remains the same, with some resolution loss, while the input image contrast changes drastically. This explains why the CNN-based approach, which relies on both features, performed worse, while the SVM ensemble model, which only depends on the CC shape, maintained its performance. Similarly to what we have tested in the diffusion domain, the SVM model could be used in other MRI sequences such as T_2 and proton-density weight images where the CC shape persists. In this cases, the CNN-based approach would fail unless a fine-tuning in the new domain was performed.

When a tumor is present, the characterization of brain structures is essential to monitoring its advance and the degree to which these structures are affected. In this case, applying QC to detect erroneous segmentations is crucial because the proposed segmentation methods are created from normal images and do not work as expected in abnormal cases such as tumor. We tested the SVM ensemble and the CNN-based models in 10 CC segmentations from 7 subjects segmented with two methods: manual and freesurfer (3 subjects were segmented by



Subject	Grade	CC Region	Figure(s)
1	IV	on	a),b)
2	IV	on	c),d)
3	IV	around	e),f)
4	III	on	g)
5	III	around	h)
6	III	around	i)
7	IV	on	j)

(k)

Figure 6: Patients with tumor depicting the contour of the CC segmentation mask: (a) Subject 1 with manual segmentation (Correct), (b) Subject 1 with Freesurfer segmentation (Incorrect), (c) Subject 2 with manual segmentation (Correct), (d) Subject 2 with Freesurfer segmentation (Incorrect), (e) Subject 3 with manual segmentation (Correct), (f) Subject 3 with Freesurfer segmentation (Incorrect), (g) Subject 4 with manual segmentation (Correct), (h) Subject 5 with manual segmentation (Correct), (i) Subject 6 with manual segmentation (Correct), (j) Subject 7 with Freesurfer segmentation (Incorrect), and (k) table displaying grade and region of the tumor for each subject.

both methods). The manual segmentation masks were correct, while the ones generated by Freesurfer were all incorrect. The classical approach had $AUC = 70.83\%$, failing to assign the correct class for 4 segmentations (Figs. 6a, 6c, 6f and 6h), when used with a typical separation threshold of 50%. This approach relies on the CC shape, which is affected by the presence of the tumor. Adapting the SVM ensemble to work in this problem is not an easy task because it is complex to model all the possible deformations of the CC being affected by a tumor using only shape features.

On the other hand, the CNN has contextual information, so it can verify not only the mask but the image. For this reason, this approach's performance had an AUC of 83.33%, with only two miss-classified segmentations (Figs. 6a and 6f). These two segmentations were incorrectly classified by the classical approach too. The tumor almost entirely invaded the CC in the first case (Fig. 6a), and several regions of the tumor had similar intensity as the CC, resulting in a false positive. In contrast, the second case (Fig. 6f) resulted in a false negative: the CC was not directly affected by the tumor, its shape was preserved, and the Freesurfer's segmentation error was subtle when compared to the other Freesurfer segmentations. It is important to point out that the CNN was not trained with tumor images, meaning the model does not know cancer

tissue, so the result could be improved if the model is fine-tuned with this kind of image.

5. Conclusions

QC is an important step intended to assure that data can be used reliably, being essential in studies with large datasets where automatic segmentation methods must be applied, and errors can be propagated and scaled faster. In medical image, QC segmentation is a neglected issue, with few proposed works. Notably, on CC segmentation, QC has been approached only on our previous proposal.

Our proposal of using the ResNet18 CNN resulted in a simple but effective tool to perform quality control over segmentation in different CC datasets. In T_1-MRI, the CNN had a similar performance as the classical machine learning framework, with an identical AUC value and requiring the same time at the testing stage. However, building the SVM-based model required much more effort due to the hand-craft feature extraction process associated with classical machine learning models. This manual process is not necessary in CNN models, where the convolutional layers learn the optimal features automatically.

Since machine learning models are data-driven, training and testing in homogeneous and normal datasets are relatively straightforward. However, trained models usually do not work well in new domains. Therefore, the extension of the CNN-based approach to domains where the image intensity changes, such as diffusion MRI, requires domain adaptation. In this work, domain adaptation was not possible because of the lack of ground-truth in the target domain; therefore the SVM-based approach got the best performance. Tumor cases are complex for any machine learning model because the target structure, the CC in this case, can vary in unexpected ways, affecting its shape, intensity, and location. However, the CNN-based approach was best suited to solve tumor cases thanks to the contextual information that it grasps from the images. Further, domain adaptation techniques, such as fine-tuning, can improve the performance of the CNN in these cases, allowing the network to learn tumor particularities.

In summary, the CNN is more versatile than the classical machine learning model. It is better able to adapt to abnormal unseen samples and is best suited to learn different problems with domain adaptation. Moreover, the same CNN architecture, using the proposed input image/mask arrange, can be used to perform QC over other brain structures.

Conflict of interest statement

We have no conflicts of interest to declare.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP - process CEPID 2013/07559-3) and by the National Counsel of Technological and Scientific Development (processes 190557/2014-1 and 308311/2016-7). We would also like to thank Prof. Heath Pardoe, Associate Professor at the Department of Neurology of the NYU Langone Medical Center, New York University, for providing us with the automatic and manually corrected segmentations from the ABIDE database.

References

[1] S. Hofer, J. Frahm, Topography of the human corpus callosum revisited—comprehensive fiber tractography using diffusion tensor magnetic resonance imaging, NeuroImage 32 (2006) 989–994.

[2] G. Cover, W. Herrera, M. Bento, S. Appenzeller, L. Rittner, Computational methods for corpus callosum segmentation on MRI: A systematic literature review, Comput. Methods and Program. in Biomed. 154 (2018) 25–35.

[3] N. Gordillo, E. Montseny, P. Sobrevilla, State of the art survey on MRI brain tumor segmentation, Magn. Reson. Imaging 31 (2013) 1426–1438.

[4] N. R. Pal, S. K. Pal, A review on image segmentation techniques, Pattern recognition 26 (1993) 1277–1294.

[5] A. Işın, C. Direkoğlu, M. Şah, Review of mri-based brain tumor image segmentation using deep learning methods, Procedia Computer Science 102 (2016) 317–324.

[6] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, et al., Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study, Journal of Cardiovascular Magnetic Resonance 21 (2019) 18.

[7] S. Bouix, M. Martin-Fernandez, L. Ungar, M. Nakamura, M.-S. Koo, R. W. McCarley, M. E. Shenton, On evaluating brain tissue classifiers without a ground truth, Neuroimage 36 (2007) 1207–1224.

[8] M. B. Abdallah, M. Blonski, S. Wantz-Mézières, Y. Gaudeau, L. Taillandier, J.-M. Moureaux, Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset, in: Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE, pp. 4403–4406.

[9] E. T. Klapwijk, F. Van De Kamp, M. Van Der Meulen, S. Peters, L. M. Wierenga, Qoala-t: A supervised-learning tool for quality control of freesurfer segmented mri data, NeuroImage 189 (2019) 116–129.

[10] V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, B. Glocker, Reverse classification accuracy: predicting segmentation performance in the absence of ground truth, IEEE transactions on medical imaging 36 (2017) 1597–1606.

[11] R. Robinson, V. V. Valindria, W. Bai, H. Suzuki, P. M. Matthews, C. Page, D. Rueckert, B. Glocker, Automatic quality control of cardiac mri segmentation in large-scale population imaging, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 720–727.

[12] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative, et al., Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control, NeuroImage 195 (2019) 11–22.

[13] W. G. Herrera, M. Pereira, M. Bento, A. T. Lapa, S. Appenzeller, L. Rittner, A framework for quality control of corpus callosum segmentation in large-scale studies, Journal of Neuroscience Methods 334 (2020) 108593.

[14] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[15] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, N. Kim, Deep learning in medical imaging: general overview, Korean journal of radiology 18 (2017) 570–584.

[16] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org.

[17] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, B. J. Erickson, Deep learning for brain mri segmentation: state of the art and future directions, Journal of digital imaging 30 (2017) 449–459.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[19] M. Schwarz, H. Schulz, S. Behnke, Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features, in: 2015 IEEE international conference on robotics and automation (ICRA), IEEE, pp. 1329–1335.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp. 248–255.

[21] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning?, IEEE transactions on medical imaging 35 (2016) 1299–1312.

[22] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (2019) 60.

[23] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, et al., Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain, Neuron 33 (2002) 341–355.

[24] R. K. Hiess, R. Alter, S. Sojoudi, B. Ardekani, R. Kuzniecky, H. Pardoe, Corpus callosum area and brain volume in autism spectrum disorder: quantitative analysis of structural mri from the abide database, Journal of autism and developmental disorders 45 (2015) 3107–3114.

[25] B. Ardekani, yuki module of the automatic registration toolbox (art) for corpus callosum segmentation, http://www.nitrc.org/projects/art (2013).

[26] S. N. Niogi, P. Mukherjee, B. D. McCandliss, Diffusion tensor imaging segmentation of white matter structures using a reproducible objective quantification scheme (roqs), Neuroimage 35 (2007) 166–174.

[27] T. E. Oliphant, A guide to NumPy, volume 1, Trelgol Publishing USA, 2006.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: NIPS-W.

[29] B. A. Ardekani, J. Kershaw, M. Braun, I. Kanuo, Automatic detection of the mid-sagittal plane in 3-d brain images, IEEE transactions on medical imaging 16 (1997) 947–952.

[30] P. Getreuer, Linear methods for image interpolation, Image Processing On Line 1 (2011) 238–259.

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[32] D. Le Bihan, J.-F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, H. Chabriat, Diffusion tensor imaging: concepts and applications, Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 13 (2001) 534–546.

[33] M. R. Nazem-Zadeh, S. Saksena, A. Babajani-Fermi, Q. Jiang, H. Soltanian-Zadeh, M. Rosenblum, T. Mikkelsen, R. Jain, Segmentation of corpus callosum using diffusion tensor imaging: validation in patients with glioblastoma, BMC Med. Imaging 12 (2012) 1.

[34] L. Rittner, J. S. Campbell, P. F. Freitas, S. Appenzeller, G. B. Pike, R. A. Lotufo, Analysis of scalar maps for the segmentation of the corpus callosum in diffusion tensor fields, J. Math. Imaging Vis. 45 (2013) 214–226.

[35] P. Freitas, L. Rittner, S. Appenzeller, R. Lotufo, Watershed-based segmentation of the midsagittal section of the corpus callosum in diffusion MRI, 24th SIBGRAPI Conf. on Graph., Patterns and Images (2011) 274–280.

[36] S. N. Niogi, P. Mukherjee, B. D. McCandliss, Diffusion tensor imaging segmentation of white matter structures using a Reproducible Objective Quantification Scheme (ROQS), NeuroImage 35 (2007) 166–174.

[37] Y. Kong, D. Wang, L. Shi, S. C. N. Hui, W. C. W. Chu, Adaptive distance metric learning for diffusion tensor image segmentation, PLoS ONE 9 (2014) 1–11.

[38] R. Luis-García, C.-F. Westin, C. Alberola-López, Gaussian mixtures on tensor fields for segmentation: Applications to medical imaging, Comput. Med. Imaging Graph. 35 (2011) 16–30.

[39] A. Omuro, L. M. DeAngelis, Glioblastoma and other malignant gliomas: a clinical review, Jama 310 (2013) 1842–1850.

On Line Sources

6 Final remarks

The automatic quality evaluation is an important component to guarantee reliability in every step of a whole image analysis pipeline. Incorrect segmentations can introduce early errors in the process. In practice, QC is performed by visual inspection on a one-to-one basis. In massive datasets, automated methods are used to produce segmentations on a large scale, and manual quality verification is unfeasible.

Although QC methods have been proposed for cardiac (VALINDRIA et al., 2017; ROBINSON et al., 2017; ROBINSON et al., 2018; ROBINSON et al., 2019), brain (ABDALLAH et al., 2016; KLAPWIJK et al., 2019; ROY et al., 2019), and multi-purpose (VALINDRIA et al., 2017) MRI applications, no specific methods for QC on CC have been formulated. Although ABDALLAH et al. (2016) studied glioma segmentation QC itself, none of these methods tested generalization of their method to cases with tumor presence. Except for ABDALLAH et al. (2016) that used FLAIR-MRI (because of the best-offered contrast for glioma), all of these methods used the T_1-MRI sequence exclusively. In this work, two methods for performing QC over CC segmentations were proposed: the first method based on classical machine learning and the second one, using a CNN. Furthermore, both methods were analyzed independently in T_1-MRI and compared in real additional scenarios: diffusion MRI sequence and patients with tumor.

The first proposed method used a handcrafted shape feature than can be extracted at several resolutions named shape signature. At one specific resolution, the shape signature was useful to evaluate segmentations in diffusion MRI using a ground truth in T_1-MRI. The ground truth in diffusion MRI is scarce and inaccurate because the diffusion images have low resolution. Furthermore, no registering process was necessary because the comparison between T_1 and diffusion was performed through the shape signature. However, this process has two important disadvantages: first, just one manually-selected resolution was used, mining the versatility of our descriptor and its capability to describe the segmentation at several levels of detail; second, although the method suppresses the need to use a ground-truth in diffusion, it still requires the ground-truth in T_1-MRI.

Classical machine learning allowed us to handle these two drawbacks using an ensemble of SVM. The selection of the proper resolutions to perform QC was done automatically by grouping similar resolutions into clusters and selecting only one resolution by each group. With the selected resolutions, an ensemble learned to distinguish correct from incorrect segmentations through supervised training. Because the ensemble learned the characteristic shape of the CC at several levels of detail, the ground-truth is no longer necessary. This framework was trained and tested in T_1-MRI normal subjects achieving an

AUC of 98.25%. The QC score given by the framework ranges between 0% and 100%, and it is independent of the decision threshold used to classify the segmentation into correct or incorrect. Two characteristics make our framework applicable in several situations: it is possible to set up the best decision threshold to minimize the false positives or false negatives occurrences, and it is independent of the image intensity associated with the segmentation mask. Independence from the image intensity allows us to use the framework in other sequences, such as T_2 or diffusion MRI. However, it becomes a disadvantage when the tested CC is considered abnormal, such as in fetal, newborn, elderly, or tumor populations, because the model does not have contextual information of the image to distinguish an abnormal segmentation from an incorrect one.

Another method, based on deep learning, to perform QC over CC segmentations was proposed using the pre-trained ResNet18 CNN. The input was arranged in a three-channel fashion: the first and the third channels were the T_1 image and the segmentation mask, respectively. The second channel was composed as the T_1 image multiplied by the segmentation mask, giving to the network explicit information about the context of the mask. The training process was alleviated using the pre-trained network, making it faster and resulting in a model less dependent on the size of the dataset. In the test dataset, the CNN achieved an AUC of 97.98%.

It is not possible to perform a direct comparison with the literature. However, two studies deserve consideration because they tested in large populations and reported similar performance as ours. ROBINSON et al. (2019) got $accuracy = 95\%$ on a cardiovascular MRI dataset composed of 4805 images using RCA. KLAPWIJK et al. (2019) got an $AUC = 98\%$ on a 784 subjects brain MRI dataset segmented with FreeSurfer using a classical machine learning approach. However, this last work only works with brain FreeSurfer segmentations, and it gives one overall score of several brain structures (not including the CC). Although these two methods got remarkable results, generalization was not tested to include other sequences or abnormal populations.

We compared the classical and the deep learning models, both of them trained in T_1 MRI data. In terms of training time, the CNN spent at least five times longer to be trained, using a GPU, than the SVM ensemble (141 and 28 minutes, respectively). For testing, both approaches employed 9 seconds approximately in a dataset with 136 samples. In terms of AUC, both approaches got a similar value ($AUC \approx 98\%$). Although the accuracy was very close, the SVM ensemble presented the best accuracy ($Accuracy \approx 92\%$) for few samples of training (below 274 images), while the CNN performed better for more available training samples (over 342 images). It makes sense for any deep learning approach where much data is necessary to achieve good performance. However, the CNN still had a reasonable performance at reduced dataset sizes, thanks to the use of pre-training and fine-tuning.

The SVM ensemble generalized better in diffusion MRI than the CNN being trained with T_1 images, because the first model is independent of the image intensity while the second one learned the intensity of the input image. In contrast, regarding of generalization in the tumor patients' dataset, the CNN got better performance because it contextualizes the image, distinguishing between a segmentation associated with an abnormal CC and an incorrect segmentation. It is not possible for the SVM ensemble where the model only considers the mask shape.

In summary, both models can be used in real applications of T_1-MRI as long as the CC has a regular shape. In other sequences, such as diffusion MRI, it can be challenging to train the model because of the lack of ground-truth. In these cases, it is recommendable to use the SVM ensemble, because this model generalizes well the learned shape to other sequences. In abnormal CC datasets, such as in fetal, newborn, elderly, or tumor populations, the CNN prevails because it gets the contextual information of the image. It is important to notice that the CNN approach is more versatile to learn new shapes and image intensities, improving its performance in domains where it is possible re-training or fine-tuning the model in the target domain.

6.1 Future work

The shape signature is a versatile descriptor capable of describing the segmentation shape at multiple levels of detail. Furthermore, since the shape signature is independent of the image intensity, it is applicable among several MRI sequences. We foresee that several populational studies could benefit from the multi-resolution shape descriptor. The CC shape descriptor could be used to characterize group differences such as sex, age, or presence of an specific pathology. While lower resolutions allow to describe global differences between subjects and therefore they are useful in discriminating subjects from different populations, higher resolutions are suitable to follow subtle changes in the structure derived from the pathology progression or the treatment effectiveness. In these cases, the selection of the proper resolutions that best describe the structure of interest is essential to solve the problem. Some applications may require a combination of different resolutions. The manual process of selection of the resolutions requires exploration of the curves being arduous and sub optimal. On the other hand, machine learning techniques can automate the selection process and ease the combination of the best resolutions. Unsupervised clustering algorithms can be valuable to group subjects, helping to find patterns or subtle differences on CC shape.

As in diffusion, in other MRI sequences such as T_2, FLAIR and proton-density is difficult to obtain a balanced training dataset composed of segmentations from different methods to train a QC model. Therefore, our SVM-based framework can be directly used

to perform QC taking advantage to the fact that the CC maintains its shape. Furthermore, more descriptors can be added to the SVM ensemble, such as area or texture, increasing its scope and gaining performance. Particularly, false negative cases of the SVM ensemble are related with missing portions of the segmentation mask that can be solved with a area preliminary verification.

Quality over other sub-cortical brain structures such as hippocampus, hypothalamus,or amygdala can be verified using either the classical or the deep approach. Although there are available some datasets with segmentation of sub-cortical structures, establishing a heterogeneous dataset with segmentations from different methods is fundamental to guarantee the algorithm generalization. In any case, the 2D slice, on which the quality assessment will be made, must be defined. For some structures, 2D characterization can be poor due to the size, orientation or physiology of the structure. In these cases, a 3D implementation of the QC method can be useful. For the classic approach, a 3D shape descriptor must be extracted and the resolution could be associated with the mesh setting. The selection and consensus of the 3D descriptors can still be made using machine learning. For the deep approach, 3D CNN classifiers are being used more and more with several pre-trained architectures available. 2.5D proposals can be considered too.

The CNN-based approach was more versatile and was able to deal better with abnormal samples. Other abnormal populations such as newborn, elderly, and atrophies, could benefit from this method. In these scenarios, including tumor, where few data are available, few-shot and zero-shot learning techniques could be employed to adapt the model to the new domain. Considerations among weight initialization, architectures, data augmentation, and meta-learning (learn how to adapt to novel classes/samples) are crucial to succeed in the task. Traditional data augmentation techniques produce only limited alternative data. Therefore, we can use generative algorithms, such as generative adversarial networks, to produce a much broader set of images increasing the model generalization. Semisupervised techniques can alleviate need for huge datasets at training stage. For example, using autoencoders for coarse weight initialization and adjusting using fine tuning can make the dataset usage more effective. Also, using self-supervised schemes, a network can learn useful representations to classify segmentations by contrastive learning. The aim is maximize the agreement between representations of the sample and its augmentations while rejecting any other sample. Our final goal is construct a framework to perform QC on several brain structures using general and customized algorithms, depending on the target, able to deal with different sequences and plausible abnormalities.

6.2 Publications

6.2.1 Relevant publications

These are the main publications of my Ph.D., covered throughout this document:

- **Corpus Callosum Shape Signature for Segmentation Evaluation**. International Federation for Medical and Biological Engineering Proceedings, 2018. Published article (HERRERA; BENTO; RITTNER, 2019).
- **A framework for quality control of corpus callosum segmentation in large-scale studies**. Journal of Neuroscience Methods, 2020. Published article (HERRERA et al., 2020).
- **Automatic quality control on corpus callosum segmentation: Comparing deep and classical machine learning approaches**. Neural Computing and Applications, 2020. Submitted article.

6.2.2 Additional publications

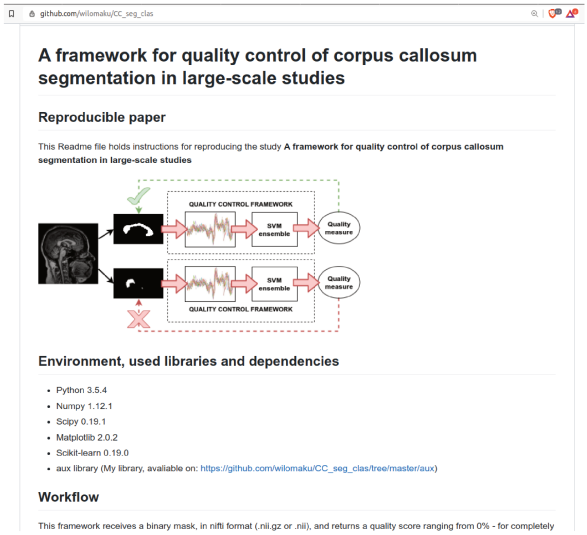
One additional study was made in collaboration with a laboratory partner applying deep learning techniques. This work was presented in the 5th Brainn Congress.

- **Classification of Alzheimer's patients and cognitive deficit through MRI**. Journal of Epilepsy and Clinical Neurophysiology, 2018. Published article (PEREIRA MARIANA; RITTNER, 2018).

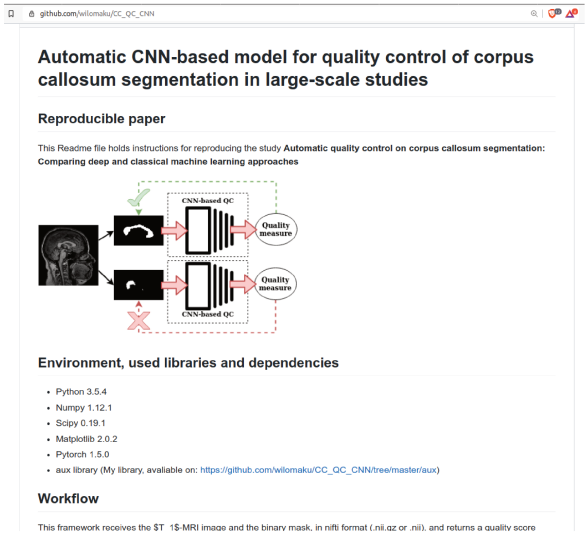
6.3 Tools

All the experiments made throughout this work are open source, and were implemented with Python (ROSSUM; DRAKE, 2009) along with specialized Python packages, among which we can mention Numpy (OLIPHANT, 2006), Scikit-learn (PEDREGOSA et al., 2011), and Pytorch (PASZKE et al., 2017). All the code, the trained models, and the instructions to use and reproduce the work are available in GitHub: Framework for QC of CC segmentation using a SVM ensemble[1] (Fig. 6.1a) and Automatic CNN-based model for QC of CC segmentation[2] (Fig. 6.1b). The data can not be openly shared due to property rights.

[1] https://github.com/wilomaku/CC_seg_clas
[2] https://github.com/wilomaku/CC_QC_CNN



(a)



(b)

Figure 6.1 – Print screen of GitHub code repositories: a) Framework for QC of CC segmentation using a SVM ensemble available at https://github.com/wilomaku/CC_seg_clas, b) Automatic CNN-based model for QC of CC segmentation available at https://github.com/wilomaku/CC_QC_CNN.

Bibliography

ABDALLAH, M. B. et al. Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset. In: IEEE. *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. [S.l.], 2016. p. 4403–4406. Cited 3 times in pages 15, 16, and 59.

ABOITIZ, F. et al. Fiber composition of the human corpus callosum. *Brain Research*, v. 598, n. 1–2, p. 143 – 153, 1992. Cited in page 23.

ADAMSON, C. et al. Software pipeline for midsagittal corpus callosum thickness profile processing: automated segmentation, manual editor, thickness profile generator, group-wise statistical comparison and results display. *Neuroinformatics*, v. 12, n. 4, p. 595–614, 2014. Cited in page 23.

BACKHAUSEN, L. L. et al. Quality control of structural mri images applied using freesurfer—a hands-on workflow to rate motion artifacts. *Frontiers in neuroscience*, v. 10, p. 558, 2016. Cited 3 times in pages 13, 14, and 15.

BAMMER, R. Basic principles of diffusion-weighted imaging. *European journal of radiology*, v. 45, n. 3, p. 169–184, 2003. Cited in page 22.

BASSER, P. J.; MATTIELLO, J.; LEBIHAN, D. MR diffusion tensor spectroscopy and imaging. *Biophysical journal*, v. 66, n. 1, p. 259–267, 1994. Cited in page 22.

BIHAN, D. L. et al. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology*, v. 161, n. 2, p. 401–407, 1986. Cited in page 22.

BIHAN, D. L. et al. Diffusion tensor imaging: concepts and applications. *Journal of magnetic resonance imaging : JMRI*, v. 13, n. 4, p. 534–546, 2001. ISSN 1053-1807. Cited in page 22.

BIHAN, D. L. et al. Measuring random microscopic motion of water in tissues with MR imaging: a cat brain study. *Journal of computer assisted tomography*, v. 15, n. 1, p. 19–25, 1991. Cited in page 22.

BLOCH, F. Nuclear induction. *Physical review*, v. 70, n. 7-8, p. 460, 1946. Cited in page 21.

BLOEMBERGEN, N.; PURCELL, E. M.; POUND, R. V. Relaxation effects in nuclear magnetic resonance absorption. *Physical review*, v. 73, n. 7, p. 679, 1948. Cited in page 21.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ACM. *Proc. of the fifth annu. workshop on Computational Learn. Theory*. [S.l.], 1992. p. 144–152. Cited in page 25.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995. ISSN 1573-0565. Cited in page 25.

COVER, G. et al. Computational methods for corpus callosum segmentation on mri: A systematic literature review. *Computer methods and programs in biomedicine*, v. 154, p. 25–35, 2018. Cited 5 times in pages 13, 14, 20, 23, and 24.

DAMADIAN, R. Tumor detection by nuclear magnetic resonance. *Science*, v. 171, n. 3976, p. 1151–1153, 1971. Cited in page 20.

DAMADIAN, R. et al. Field focusing nuclear magnetic resonance (FONAR): visualization of a tumor in a live animal. *Science*, v. 194, n. 4272, p. 1430–1432, 1976. Cited in page 20.

DICE, L. R. Measures of the amount of ecologic association between species. *Ecology*, v. 26, n. 3, p. 297–302, 1945. Cited in page 23.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.: s.n.], 2012. Cited in page 25.

EDELMAN, R. R.; WARACH, S. Magnetic resonance imaging. *N. Engl. J. Med.*, v. 328, n. 10, p. 708–716, 1993. Cited 2 times in pages 13 and 20.

ENDERLE, J. D.; BRONZINO, J. D. *Introduction to biomedical engineering*. [S.l.]: Academic press, 2012. Cited 2 times in pages 19 and 21.

FREITAS, P. F. Segmentação e parcelamento do corpo caloso em imagens de tensor de difusão. *Universidade Estadual de Campinas, UNICAMP, Brasil*, Mestrado em Engenharia Elétrica, 2012. Cited in page 20.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.: s.n.], 2016. http://www.deeplearningbook.org. Cited in page 26.

GORDILLO, N.; MONTSENY, E.; SOBREVILLA, P. State of the art survey on MRI brain tumor segmentation. *Magn. Reson. Imaging*, v. 31, n. 8, p. 1426–1438, 2013. Cited in page 23.

GRIMM, O. et al. Amygdalar and hippocampal volume: a comparison between manual segmentation, freesurfer and vbm. *Journal of neuroscience methods*, v. 253, p. 254–261, 2015. Cited 2 times in pages 14 and 24.

GUENETTE, J. P. et al. Automated versus manual segmentation of brain region volumes in former football players. *Neuroimage: clinical*, v. 18, p. 888–896, 2018. Cited 3 times in pages 13, 14, and 15.

HE, Q. et al. A context-sensitive active contour for 2D corpus callosum segmentation. *Int. J. of Biomed. Imaging*, v. 2007, 2007. Cited 2 times in pages 13 and 23.

HERRERA, W. G.; BENTO, M.; RITTNER, L. Corpus callosum shape signature for segmentation evaluation. In: *XXVI Brazilian Congress on Biomedical Engineering*. Singapore: [s.n.], 2019. p. 143–147. Cited 2 times in pages 27 and 63.

HERRERA, W. G. et al. A framework for quality control of corpus callosum segmentation in large-scale studies. *Journal of Neuroscience Methods*, v. 334, p. 108593, 2020. Cited 2 times in pages 33 and 63.

HOFER, S.; FRAHM, J. Topography of the human corpus callosum revisited—comprehensive fiber tractography using diffusion tensor magnetic resonance imaging. *NeuroImage*, v. 32, n. 3, p. 989–994, 2006. Cited 2 times in pages 13 and 19.

HUANG, C.; WU, Q.; MENG, F. Qualitynet: Segmentation quality evaluation with deep convolutional networks. In: IEEE. *Visual Communications and Image Processing (VCIP), 2016*. [S.l.], 2016. p. 1–4. Cited 2 times in pages 15 and 16.

HUISMAN, T. A. G. M. Diffusion-weighted imaging: basic concepts and application in cerebral stroke and head trauma. *European Radiology*, v. 13, n. 10, p. 2283–2297, 2003. Cited in page 22.

JR, C. R. J. et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, v. 27, n. 4, p. 685–691, 2008. Cited in page 13.

KAMNITSAS, K. et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, v. 36, p. 61–78, 2017. Cited in page 15.

KESHAVAN, A. et al. Mindcontrol: A web application for brain segmentation quality control. *NeuroImage*, v. 170, p. 365–372, 2018. Cited 2 times in pages 14 and 15.

KIESOW, H. et al. 10,000 social brains: Sex differentiation in human brain anatomy. *Science advances*, v. 6, n. 12, p. 1170, 2020. Cited in page 13.

KLAPWIJK, E. T. et al. Qoala-t: A supervised-learning tool for quality control of freesurfer segmented MRI data. *NeuroImage*, v. 189, p. 116–129, 2019. Cited 4 times in pages 14, 16, 59, and 60.

KONG, Y. et al. Adaptive distance metric learning for diffusion tensor image segmentation. *PLoS ONE*, v. 9, n. 3, p. 1–11, 2014. Cited in page 24.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, v. 160, p. 3–24, 2007. Cited in page 24.

LAUTERBUR, P. C. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature*, v. 242, p. 190–191, 1973. Cited in page 20.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, v. 521, n. 7553, p. 436–444, 2015. Cited in page 25.

LEE, J.-G. et al. Deep learning in medical imaging: general overview. *Korean journal of radiology*, v. 18, n. 4, p. 570–584, 2017. Cited in page 26.

LITJENS, G. et al. A survey on deep learning in medical image analysis. *Medical image analysis*, v. 42, p. 60–88, 2017. Cited in page 26.

MAKROPOULOS, A. et al. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage*, v. 173, p. 88–112, 2018. Cited 2 times in pages 14 and 15.

MOGALI, J. K. et al. A shape-template based two-stage corpus callosum segmentation technique for sagittal plane T1-weighted brain magnetic resonance images. In: IEEE. *2013 IEEE Int. Conf. on Image Process.* [S.l.], 2013. p. 1177–1181. Cited in page 23.

NAZEM-ZADEH, M. R. et al. Segmentation of corpus callosum using diffusion tensor imaging: validation in patients with glioblastoma. *BMC Med. Imaging*, v. 12, n. 1, p. 1, 2012. Cited in page 24.

OLIPHANT, T. E. *A guide to NumPy*. [S.l.: s.n.], 2006. v. 1. Cited in page 63.

PASZKE, A. et al. Automatic differentiation in PyTorch. In: *NIPS-W*. [S.l.: s.n.], 2017. Cited in page 63.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *J. of Mach. Learn. Res.*, v. 12, p. 2825–2830, 2011. Cited in page 63.

PENG, B. et al. Evaluation of segmentation quality via adaptive composition of reference segmentations. *IEEE transactions on pattern analysis and machine intelligence*, v. 39, n. 10, p. 1929–1941, 2017. Cited 2 times in pages 15 and 16.

PEREIRA MARIANA, H. W. L. R.; RITTNER, L. Classification of Alzheimer's patients and cognitive deficit through MRI. *Journal of Epilepsy and Clinical Neurophysiology*, v. 24, n. 5, 2018. Cited in page 63.

PIERPAOLI, C.; BASSER, P. J. Toward a quantitative assessment of diffusion anisotropy. *Magnetic resonance in Medicine*, v. 36, n. 6, p. 893–906, 1996. Cited in page 22.

RABI, I. I. et al. The molecular beam resonance method for measuring nuclear magnetic moments. *Physical review*, v. 55, n. 6, p. 526, 1939. Cited in page 20.

RAWAT, W.; WANG, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, v. 29, n. 9, p. 2352–2449, 2017. Cited in page 26.

REEVES, A. P.; LIU, S.; XIE, Y. Image segmentation evaluation for very-large datasets. *Proc. SPIE*, v. 9785, p. 9785–11, 2016. Cited in page 14.

REVETT, K. An introduction to magnetic resonance imaging: From image acquisition to clinical diagnosis. In: *Innovations in Intelligent Image Analysis*. [S.l.: s.n.], 2011. p. 127–161. Cited in page 21.

ROBINSON, R. et al. Real-time prediction of segmentation quality. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2018. p. 578–585. Cited 3 times in pages 15, 16, and 59.

ROBINSON, R. et al. Automatic quality control of cardiac mri segmentation in large-scale population imaging. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2017. p. 720–727. Cited 2 times in pages 15 and 59.

ROBINSON, R. et al. Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study. *Journal of Cardiovascular Magnetic Resonance*, v. 21, n. 1, p. 18, 2019. Cited 3 times in pages 15, 59, and 60.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: [s.n.], 2009. ISBN 1441412697. Cited in page 63.

ROY, A. G. et al. Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, v. 195, p. 11–22, 2019. Cited 3 times in pages 15, 16, and 59.

SCHOEMAKER, D. et al. Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *NeuroImage*, v. 129, p. 1–14, 2016. Cited in page 13.

SHI, R. et al. Visual quality evaluation of image object segmentation: subjective assessment and objective measure. *IEEE Transactions on Image Processing*, v. 24, n. 12, p. 5033–5045, 2015. Cited 2 times in pages 15 and 16.

SHI, W.; MENG, F.; WU, Q. Segmentation quality evaluation based on multi-scale convolutional neural networks. In: IEEE. *Visual Communications and Image Processing (VCIP), 2017 IEEE*. [S.l.], 2017. p. 1–4. Cited 2 times in pages 15 and 16.

STEJSKAL, E. O.; TANNER, J. E. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The journal of chemical physics*, v. 42, n. 1, p. 288–292, 1965. Cited in page 22.

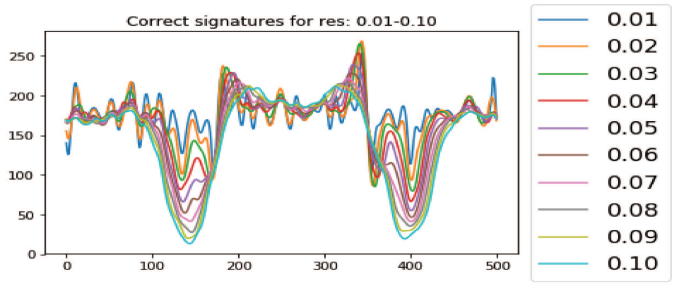
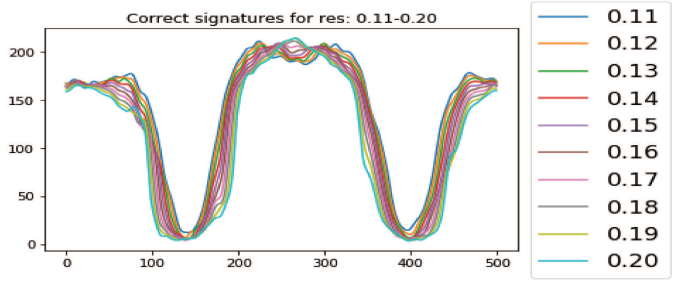
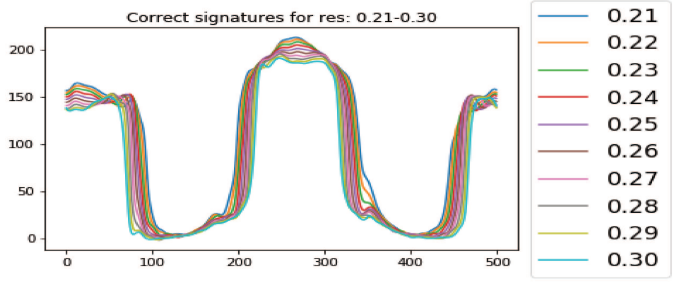
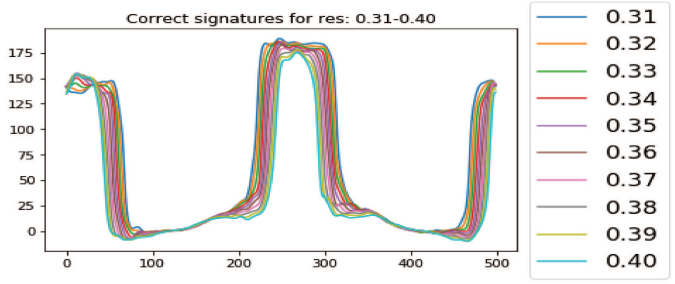
THOMPSON, P. M. et al. Enigma and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Translational psychiatry*, v. 10, n. 1, p. 1–28, 2020. Cited in page 13.

VALINDRIA, V. V. et al. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging*, v. 36, n. 8, p. 1597–1606, 2017. Cited 3 times in pages 15, 16, and 59.

WENGER, E. et al. Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains. *Human brain mapping*, v. 35, n. 8, p. 4236–4248, 2014. Cited in page 13.

Appendix

Configuration for individual SVM classifiers

SVM/*ext*	Kernel	C	Signatures
1/0.01	rbf	10	
2/0.02	rbf	20	
3/0.03	rbf	20	
4/0.04	rbf	100	
5/0.05	rbf	100	
6/0.06	rbf	100	
7/0.07	rbf	20	
8/0.08	rbf	100	
9/0.09	rbf	100	
10/0.10	rbf	100	
11/0.11	rbf	100	
12/0.12	rbf	20	
13/0.13	rbf	100	
14/0.14	rbf	100	
15/0.15	rbf	100	
16/0.16	rbf	100	
17/0.17	linear	100	
18/0.18	rbf	100	
19/0.19	rbf	100	
20/0.20	rbf	100	
21/0.21	poly	10	
22/0.22	poly	10	
23/0.23	poly	50	
24/0.24	linear	50	
25/0.25	poly	100	
26/0.26	poly	20	
27/0.27	poly	20	
28/0.28	rbf	100	
29/0.29	poly	100	
30/0.30	linear	20	
31/0.31	rbf	100	
32/0.32	poly	20	
33/0.33	poly	50	
34/0.34	poly	50	
35/0.35	rbf	100	
36/0.36	linear	0.1	
37/0.37	rbf	10	
38/0.38	rbf	10	
39/0.39	poly	100	
40/0.40	rbf	20	

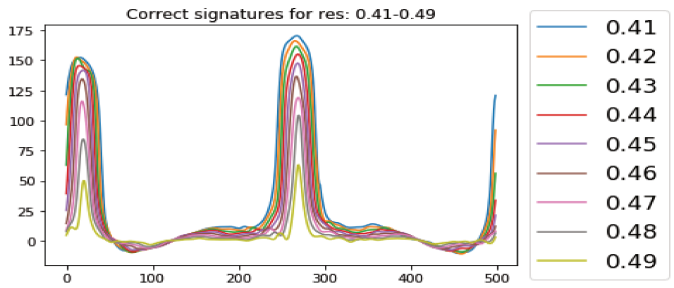
SVM/*ext*	Kernel	C	Signatures
41/0.41	linear	0.1	
42/0.42	rbf	10	
43/0.43	poly	50	
44/0.44	rbf	20	
45/0.45	linear	20	
46/0.46	rbf	50	
47/0.47	linear	100	
48/0.48	poly	50	
49/0.49	rbf	10	

Table .1 – Configuration of all SVM-resolution classifiers portraying an example signature associated with a correct segmentation for each resolution, where C is the penalty parameter and gamma is the coefficient of the rbf kernel