Fall 2013

Machine-learning based automated segmentation tool development for large-scale multicenter MRI data analysis

Eun Young Kim
University of Iowa

MACHINE-LEARNING BASED AUTOMATED SEGMENTATION TOOL DEVELOPMENT FOR LARGE-SCALE MULTICENTER MRI DATA ANALYSIS

by

Eun Young Kim

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

December 2013

Thesis Supervisor: Assistant Professor Hans J. Johnson

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Eun Young Kim

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biomedical Engineering at the December 2013 graduation.

Thesis Committee:

Hans J. Johnson, Thesis Supervisor

Joseph M. Reinhardt

Vincent A. Magnotta

Michael A. Mackey

Mona K. Garvin

ACKNOWLEDGEMENTS

To my advisor, Hans J. Johnson for his guidance through the project. To my colleague and friend, Joy, for all the days and nights we have spent in the hospital helping each other. To the staff in the SINAPSE lab for their assistance and for their collaboration. To PREDICT-HD Study group for this opportunity to learn and research. To my roommate, Parang, for all the years of sharing the long journey. To my new roommate, Yujaung, for all the encouragement. To Heejin Ohn, for all his patience and love. Last but not the least, to my parents and family, KwangJoon Kim, Misun Ohm, and Dongjin Kim, for their unconditional support and love all theses years.

Thank you.

ABSTRACT

Background: Volumetric analysis of brain structures from structural Magnetic Resonance (MR) images advances the understanding of the brain by providing means to study brain morphometric changes quantitatively along aging, development, and disease status. Due to the recent increased emphasis on large-scale multicenter brain MR study design, the demand for an automated brain MRI processing tool has increased as well. This dissertation describes an automatic segmentation framework for brain subcortical structures that is robust for a wide variety of MR data.

Method: The proposed segmentation framework, *BRAINSCut*, is an integration of robust data standardization techniques and machine-learning approaches. First, a robust multi-modal pre-processing tool for automated registration, bias correction, and tissue classification, has been implemented for large-scale heterogeneous multi-site longitudinal MR data analysis. The segmentation framework was then constructed to achieve robustness for large-scale data via the following comparative experiments: **1)** Find the best machine-learning algorithm among several available approaches in the field. **2)** Find an efficient intensity normalization technique for the proposed region-specific localized normalization with a choice of robust statistics. **3)** Find high quality features that best characterize the MR brain subcortical structures. Our tool is built upon 32 handpicked multi-modal muticenter MR images with manual traces of six subcortical structures (nucleus accumben, caudate nucleus, globus pallidus, putamen, thalamus, and hippocampus) from three experts.

A fundamental task associated with brain MR image segmentation for research and clinical trials is the validation of segmentation accuracy. This dissertation evaluated the proposed segmentation framework in terms of validity and reliability. Three groups of data were employed for the various evaluation aspects: 1) traveling human phantom data for the multicenter reliability, 2) a set of repeated scans for the measurement stability across various disease statuses, and 3) a large-scale data from Huntington's disease (HD) study for software robustness as well as segmentation accuracy.

Result: Segmentation accuracy of six subcortical structures was improved with 1) the bias-corrected inputs, 2) the two region-specific intensity normalization strategies and 3) the random forest machine-learning algorithm with the selected feature-enhanced image. The analysis of traveling human phantom data showed no center-specific bias in volume measurements from *BRAINSCut*. The repeated measure reliability of the most of structures also displayed no specific association to disease progression except for caudate nucleus from the group of high risk for HD. The constructed segmentation framework was successfully applied on multicenter MR data from PREDICT-HD [127] study ($< 10\%$ failure rate over 3000 scan sessions processed).

Conclusion: Random-forest based segmentation method is effective and robust to large-scale multicenter data variation, especially with a proper choice of the intensity normalization techniques. Benefits of proper normalization approaches are more apparent compared to the custom set of feature-enhanced images for the accuracy and robustness of the segmentation tool. *BRAINSCut* effectively produced subcortical volumetric measurements that are robust to center and disease status with validity confirmed by human experts and low failure rate from large-scale multicenter MR data. Sample size estimation, which is crutial for designing efficient clinical and research trials, is provided based on our experiments for six subcortical structures.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ALGORITHMS

Algorithm

CHAPTER 1
BACKGROUND AND SIGNIFICANCE

Neuroimaging studies have become increasingly used to improve neuroanatomical knowledge. This chapter initiates a discussion about the importance of volumetric studies from brain magnetic resonance images (MRI) in conjunction with in-vivo research and clinical trials. The structural-MRI has long been used to reveal a relation between subcortical volumes and disease progression. Due to the recent increased emphasis on large-scale multicenter brain MR study design, the demand for an effective automated segmentation tool has been increased as well. We reviewed fundamental bases for automated segmentation tools especially for large scale data collected at different sites including literature in the field, theoretical background, and previous works with in-depth discussion.

1.1 Introduction

Robust and precise delineation of subcortical regions from large-scale multicenter brain structural MRI effectively increases knowledge of brain volumetric changes along aging, development, or disease status. A rich set of MRI-based soft tissue information provides opportunities for neuroanatomic projects detecting subtle morphological changes along disease progression quantitatively [139]. With those advanced MRI techniques, it is now well-known that trajectories of neuroanatomical morphology in normal aging differ from neurological diseases such as schizophrenia, Alzheimer's disease, autism, Huntington's disease (HD), and others [153, 66, 163, 10, 43, 11, 113, 40, 9, 115, 52, 107, 149, 114, 45, 12, 108, 13]. Recently, the emphasis on data-sharing and translational research for clinical trials has lead to the collection of multiple large-scale, multicenter, and longitudinal medical imaging data sets [127, 161]. Processing such data with an efficient and reliable delineation method is critical to expand our understanding of the brain along with clinical trials.

Difficulties, however, exist in conducting robust quantitative assessments on large-scale multicenter MRI data. While manual tracing remains common practice for anatomical delineation, manual solutions are limited by the high cost of manual labor and low intra- and inter-rater consistency. This is especially true when it comes to large-scale longitudinal multicenter studies, where a manual approach presents major issues to an efficient quantitative MRI analysis. While longitudinal multicenter study design allows us to have valuable data for populations of interests (especially in rare diseases), processing such data remains a great challenge.

Several automated tools have been developed [14] to address these drawbacks to manual tracing. These tools, however, are also limited by multiple challenges. The first challenge involves the limitation of automated segmentation tools to handle large data variation, especially given that prior information is limited. An example of this is that human experts are able to delineate MRIs with dynamic inference based on empirical knowledge and pre-attained information from biology, physiology, neurology, and anatomy. However, the automated computer algorithm is constrained by restricted static information – usually voxel intensity and location. The second main challenge is associated to the fact that an automated method is not generalizable compared to its level of human experts therefore providing consistent information to the algorithm is crucial. However intra-scan intensity inhomogeneity, one of the inevitable obstacles when using MRI, poses a primary challenge for providing full knowledge with strong consistency for the successful automated segmentation. Third, for large-scale multicenter MRI data, the intensity profile variability is even further heterogeneous compared to the intensity profile from traditional single site studies. These multicenter variations arise from diversities in manufacturers, protocol, and field strengths of MR scanners. Fourth, time and computation resources are often limited to test and apply against large-scale data, especially when multiple iterations are required between testing and development. Due to these four main challenges, extracting reliable volumetric information from MRIs is still an open research area.

There have been few attempts to develop an automated framework for processing large-scale multicenter MR data. Among the various available methodologies, the machine-learning based segmentation algorithm is attractive since generalizable and robust techniques for scalable data analysis is provided. Unfortunately, there is little research that provides a broad and conclusive study showing the relative benefits of each of the components in the cumulative framework.

This work unifies multiple image processing methods under a common segmentation framework using appropriate criteria. In this chapter, we provide a review on the relative literature including subcortical segmentation and associated pipeline methods in the field (Section 1.2), a basis of this study with regard to structural MRI principles in clinical research (Section 1.2.2), related challenges (Section 1.3), and preliminary works with in depth discussion (Section 1.6).

1.2 Literature Review

This section reviews the established works that are highly related to our automated segmentation framework development. We review the studies about the subcortical regions from brain MRI in chronological order (Section 1.2.1) [1], which includes also large-scale MRI data analysis (Section 1.2.1.2). Note that the scope of this review is limited to MRI data analysis that is most relevant to large-scale multicenter neuroimaging studies.

1.2.1 Subcortical Volumetry Study

The relation between subcortical volumetry, obtained from structural-MRI, and neuronal disease has been studied extensively over the last two decades, yet there continues to be an active and productive community contributing enhancements.

One of the early studies of the subcortical regions from MR images was initiated by proposing manual protocols for hippocampus and amygdala volumes from 11 healthy adults [171]. Later, manual volumetric measurements in computerized systems were obtained from 15 patients with mild Huntington's disease and 19 age- and gender-matched control subjects. The conclusion was that putamen is a more sensitive indicator of brain abnormalities in mild HD than measures of caudate atrophy [66]. The study [153] also confirmed that subcortical atrophy is significantly correlated

[1] This review literature is searched from PubMed library (`http://www.ncbi.nlm.nih.gov/pubmed/`) by using keyword of *'subcortical segmentation'* and excluded for animal studies, lesion segmentation studies, and functional MRI studies.

with specific cognitive deficits in HD, and demonstrated that cortical atrophy also has an important association with the cognitive deficits of patients with Huntington's disease. Atrophy in hippocampal and cortical area in subcortical ischemic vascular dementia (SIVD) [46] and different annual rate of hippocampus atrophy [22] were also studied using brain MRIs.

Along with these projects, a large number of subcortical volumetric studies have been introduced in the field. The next two sections review the automated segmentation and subcortical volumetric study (Chapter 1.2.1.1) and the sub-category that are specific to the large-scale study in the last decade (Chapter 1.2.1.2)

1.2.1.1 Automated Segmentation and Subcortical Volumetric Study

Over the past decade, questions have been raised regarding subcortical volume characteristics, especially along the aging or disease progression. Such diseases of interest often include neurodegenerative diseases (i.e., Alzheimer, Parkinsons disease, and Huntingtons disease), as well as other neuronal diseases (i.e., Schizophrenia, multiple sclerosis, Downs syndrome) (See Table 1.1). In early studies (∼2009), researchers primarily focused on the development and validation of automatic subcortical segmentation methodologies. Encouraged by those methodology accomplishments, subcortical volumetric analysis between different populations came along (∼2010). Recently (2011∼2013), researchers have focused their attention on larger sample studies, i.e. $n > 100$, to increase power of the comparative study and its findings.

As shown in Table 1.1, subcortical brain structures often investigated independently or together by biological clusters, i.e., subcortical or striatal volume as shown in Figure 1.1. In either cases, the most common findings are the volumetric differences between disease and normal healthy control populations, i.e., lower or a decrease in subcortical volumes in comparison to normal healthy controls. In contrast, there were a number of studies that reported no significant differences between disease and normal populations [79, 23, 154, 59, 55, 158, 167].

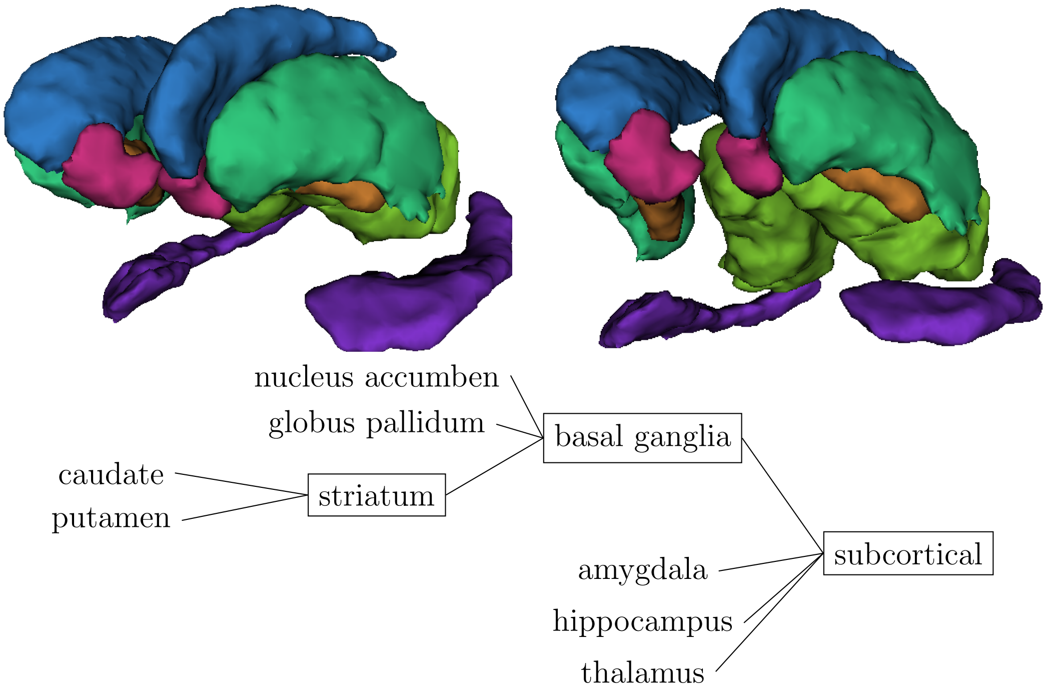


Figure 1.1: Subcortical structure of hierarchical definition. Subcortical structure often includes seven main structures of 1) nucleus accumben (magenta), 2) globus pallidum (orange), 3) caudate (blue), 4) putamen (green), 5) amygdala, 6) hippocampus (purple), and 7) thalamus (avocado green). In this study, we focuses on the segmentation tool development for all the subcortical structures but amygdala.

Table 1.1: Literature review on subcortical segmentation

Case	Nuc. Acc	Amygdala	Caudate	Glob. Pal.	Hippocampus	Putamen	Thalamus	Sc/Bg/St
Dvlpmnt	↑ [109]	↑ [109]	↑ [109]	↑ [109]	↑ [109]	↑ [109]		
Aging	↓ [93]			↓ [93]	↑ [93]	↓ [93]	↓ [93]	
Alzhmr					↓ [174]	↓ [34]	↓ [174, 34]	↓ Sc [174]
HD				↓ [40]			↓ [40]	↓ St [40]
-preHD	↓ [95, 165]	↓ [95]	↓ [95, 95, 165]	↓ [95]	↓ [95]	↓ [95, 95, 165]		↓ St [174]
MS	↓ [130]			↓ [130]	↓ [39]	↓ [130]	↓ [130]	↓ Sc [144]
Prknsn				↓ [99]	↓ [99]	↓ [99]	↓ [99]	↓ Sc [17], St [159]
Psych.	↓ depression [87]	↓ anxiety [18]	↓ PME [36]		↓ depression [162, 87]		↓ depression [98]	
cond.	↑ apathy [112]	↑ suicidal [151]	↓ Schz [24]		↓ Schz [55]		↓ Schz [24]	
Etc.			↓ McLd [164]		↓ ME [98]	↓ McLd [164]	↓ ME [98]	↓ ME [Sc][98]
					↓ MCI & SVD [42]			
					↓ OSA [160]			
		↑ HIV [26]	↓ HIV [16]	↑ Divers [98]	↑ memory [67, 35, 35]	↓ HIV [16]	↑ Divers [98]	
		↑ EO [173]			↑ EO [173]			
					⊙ WM lesion [42]			⊙ WM lesion [5]
		⊙ FTD [132]						

Literature review on subcortical segmentation reporting either increased (denoted ↑) or decreased (denoted ↓) volume in relation with the specified disease/condition. Most instances reported atrophy or lesser volume with diseased population as well as aging. Only hippocampus showed incrased volume with aging. (**Acronym:** Sc=Subcortical, Bg=Basal ganglia, St=Striatum, $Dvlpmnt$=Development, $Alxhmr$=Alzheimer, ME=Myoclonic Epliepsy, SVD=subcortical vascular dementa, OSA=obstructive sleep apnea, $McLd$=McLeod Syndrom, $Schz$=Schizophrenia, EO=Elderly Obess)

1.2.1.2 Scalable MRI Data Analysis Study

Here we review a few studies that were applied on relatively large-scale MRI data set, which shows the recent study trend of research scale. The study by Zijdenbos et al. [182] applied an automatic pipeline on a number of large-scale multi-center study data (n > 1000 scans) for studying multiple sclerosis. Another multicenter study (n = 103) [106] concluded that hippocampal volume was the primary determinant of memory decline, whereas executive function decline was related to multiple brain components. Analysis of 71 subjects with pre-symptomatic to advance HD in the study [25] confirmed that gray-matter and white-matter volumes are significantly smaller than healthy controls ($n = 24$).

There are more of Huntington's disease (HD) related projects that employ multicenter MRI data, empowered by PREDICT-HD [127] and TRACK-HD [156] projects collecting large-scale multicenter MR data. One study [156] performed blinded analyses on the baseline cross-sectional data from 366 individuals: 123 controls, 120 premanifest (pre-HD) individuals, and 123 patients with early HD. In study [13], volumes of striatum and white matter were obtained for 170 subjects and found significantly smaller volumes in individuals who would be diagnosed 1 to 4 years following the initial MRI scan, compared with those who would remain in the pre-HD stage. In that same study, they also concluded that putamen volume was the measure that best distinguished between the two groups [13]. The study of larger subjects, $n = 657$ at 32 centers, reported that volumes of all three HD subgroups differed significantly from controls for total brain tissue, cerebralspinal fluid, white-matter, cortical gray matter, thalamus, caudate, and putamen[114]. The study also suggested that total striatum volume demonstrated the largest differences between Controls and all three prodromal subgroups [114].

We also summarized the recent (2010∼2013) projects of multicenter MRI data analysis as shown in Table 1.2 reporting number of subjects (or scans). In shorts, the recent study design trends has increased number of involved scans by using multi-protocols, multi-scanner, and/or multi-site collection. It can be easily anticipated that the demand of a automated segmentation tool has been increased with those studies with large-scale structural-MRI data, that possibly collected with non-uniform platforms.

1.2.2 Structural MRI Principles

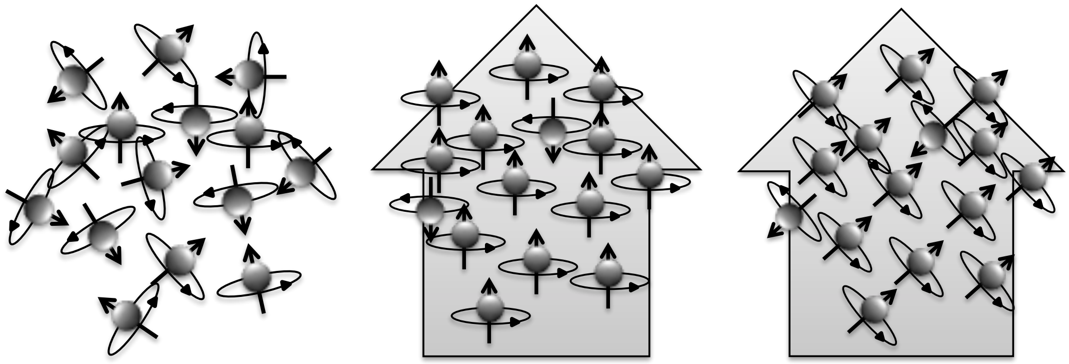
Magnetic resonance imaging (MRI) techniques allow us to see details of brain structures in-vivo non-invasively. This description of MRI summarizes two main sources: 1) ‘Principle of imaging in neuro ophthalmology’ by Rubin et al. [139] and 2) ‘The basics of MRI’ by Hornak [69].

Magnetic resonance (MR) phenomenon was discovered by Felix Bloch and Edward Purcell independently in 1946. The MRI is based on the concept of nuclear magnetic resonance, in which recording the absorption and emission of energy when the nuclei of certain atoms becomes aligned or polarized under a strong magnetic field. MRI techniques for human scanning are almost exclusive to measuring hydrogen atom, which is commonly found in body water. Randomly oriented hydrogen

Table 1.2: Recent Large Scale Study: number of subjects (scans) analyzed in the brain volumetry study from structural-MRI between 2010 and 2013.

Case	n	group n where applicable
Jochemsen et al. (2013)	663	
Becker et al. (2011)	155 men	84 vs 71
Cerasa et al. (2011)	183	74 vs 109
Dalaker et al. (2011)	84	43 vs 41
de Boer et al. (2011)	974 elderly subjects	
Eckerstrom et al. (2011)	166	92 vs 40 vs 44
Messina et al.(2011)	165	72 vs 32 vs 15 vs 46
Muller et al. (2011)	1232	- 663 follow-up
Smith et al. (2011)	145	40 vs 94 vs 11
Spoletini et al. (2011)	100	50 vs 50
Widya et al. (2011)	471	
Dewey et al. (2010)	120	
Appelman et al. (2010)	840	
Schwartz et al. (2010)	105	61 vs 44
Sabuncu et al. (2010)	282	
Geerlings et al. (2010)	1044	

atoms of the body in a natural state become parallel or anti-parallel to the magnetic (longitudinal) direction under a strong magnetic field. A brief introduction of radio frequency (RF) causes energy transfer to the proton, which results in decrease in the longitudinal magnetization and increase in transversal magnetization. After the brief exposure of RF pulse, a longitudinal magnetization increases and a transversal magnetization decreases or disappears (dephase). The different stages of magnetic resonance behavior of protons is shown in the Figure 1.2.



(a) Free protons (b) Protons with B_0 (c) Protons with B_0 and RF

Figure 1.2: Magnetic resonance behavior of protons. In the absence of an external magnetic field B_0, the spin orientation of free protons is random 1.2(a). In a strong magnetic field B_0, the free protons become aligned with their magnetic axis parallel (or, less often, antiparallel) to the magnetic field 1.2(b). Exposure to a brief radiofrequency pulse (RF) at the Larmor frequency changes the alignment of the free protons' spin axes. After the radiofrequency pulse, the free protons twirl like tops around the lines of force of the magnetic field with a motion called precession 1.2(c). Description is mostly adapted from [139, 120].

The longitudinal relaxation is described by a time constant $T1$ and the transversal relaxation is described by a time constant $T2$. $T1$ signal is sensitive to tissue composition, structure, and surroundings and $T2$ signal is sensitive to the inhomogeneity of the external and internal magnetic field. $T1$ and $T2$ weighted images provide orthogonal information for different tissue types of body (See Table 1.3), that allows us to differentiate boundaries between biologically diverse tissues.

Table 1.3: $T1$ and $T2$ signal characteristics

	$T1$ *signal*	$T2$ *signal*
Air ,Bone	Dark	Dark
High protein	Bright	Dark
Fat	Bright	Dark
Water, Cerebrospinal fluid	Dark	Bright
Very viscous protein, Dura mater	Dark	Dark
Muscle, Nerve	Dark	Dark Gray
Gray matter	Dark Gray	Light Gray
White matter	Light Gray	Dark Gray

$T1$- and $T2$-weighted MR images provide complementary intensity information each other about different tissue types in the brain image [139] and example image is also shown in Figure 1.3. This table describes different signal characteristics in $T1$- and $T2$-weighted images for eight categories of tissue types. For instances, ambiguity between gray and white matter in $T1$-weighted image is often cleared out with $T2$-weighted image together since voxels of relatively dark gray in $T1$ with light gray in $T2$ is likely to be gray matter than white matter.

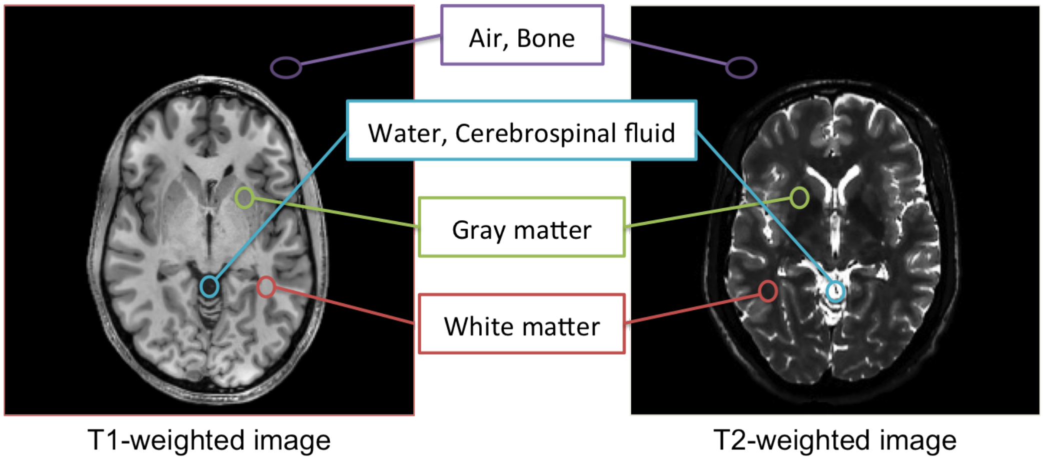


Figure 1.3: $T1$- and $T2$-weighted MR images that present distinguished characteristics for an identical tissue types, including 1) air and bone, 2) water and cerebrospinal fluid, 3) grey matter, and 4) white matter. Descriptive characteristics are presented in Table 1.3. This complimentary intensity profile between $T1$- and $T2$-weighted MRI for a certain tissue type often allows to better identify correct tissue types.

1.2.3 Huntington's Disease and structural MRIs

Huntington's disease is an autosomal dominant, monogenic neurodegenerative disorder clinically characterized by progressive involuntary movements, neuropsychiatric disturbances, and cognitive impairments [127]. At genetic level, the disease is caused by an extended trinucleotide (CAG) repeat on chromosome 4, which results in widespread neuronal degeneration preferentially within the stratum [101] (See Figure 1.1).

While there is currently no cure for Huntington's disease, structural MRIs have provided a non-invasive mean to identify important pathological brain changes from Huntington's disease. These measurable brain changes are proposed as biomarkers for many clinical trials that aim to address the devastating symptoms of HD. The most common finding in Huntington's disease from structural-MRI is subcortical atrophy. Harris et al. [66] manually identified the putamen and caudate of 34 subjects (15 patients and 19 controls) and reported greater atrophy in putamen than caudate. Another study by Starkstein et al. [153] also reported significant correlation between the CAG repeat and atrophy rates and smaller left-sided subcortical volume in Huntington's disease. Montoya et al. [101] found a more significant relationship between putamen atrophy with neurologic examination scores than caudate. Several of the most robust detectable cross-sectional change measurements in the PREDICT-HD study have shown the volumetric changes of caudate and putamen as well [11, 115, 114, 12].

1.2.4 Recent Study Trend: Multicenter Study with example of PREDICT-HD Project

Multicentral study design has rapidly gained its popularity due to several reasons. Large clinical study conducted cooperatively at multiple centers generally known to have number of benefits over a single-center study. Utilizing a multicenter approach in observational clinical study allows for improved generalizability of the results, a larger sample size, and, consequently, improved efficiency [152]. In the case of a rare disease, it is crucial to have multi-center study design to have sufficient sample size to test hypothesis.

One of example for multicentral study is the PREDICT-HD study, which is our main study focus, that collected at multicenter for Huntington's disease (HD) study. The PREDICT-HD (R01NS04006) study is an international 32-center observational study of longitudinal neuro-degeneration of persons at-risk for HD (prodromal stage) with continuous funding from 2001 to 2014. PREDICT-HD is part of a world-wide effort to provide treatments for HD, both symptomatic and prodromal. The PREDICT-HD cohort and database have become international resources and offer an unprecedented opportunity to examine the pathophysiology and neurobiology of early HD. The PREDICT-HD study is designed to identify the earliest detectable changes in cognitive skills, emotions, and anatomy as a person transitions from healthy to being diagnosed with Huntington's disease[127]. The study participants include prodromal Huntington's disease subjects and controls. The prodromal subjects are individuals who have been identified as carrying the Huntington's disease gene mutation but did not have a clinical diagnosis of Huntington's disease.

Early studies have shown differences between the prodromal and control in basal ganglia, cerebral white matter, and cerebral cortex size [113] and morphology [107]. Their findings suggest the possibility of abnormal neural development in the prodromal Huntington's disease population even before the clinically defined on-set of the disease[113, 107]. Later studies on the PREDICT-HD dataset presented consistent volumetric decrease of subcortical structures in both small and large scale data studies. The study conducted by Aylward et al. [10] showed faster change rates of the subcortical structures. Nopoulos et al [108] reported smaller intracranial volume in prodromal Huntington's disease subjects. In summary, the findings from the PREDICT-HD study using structural-MRI are very consistent, demonstrating measurable atrophy of broad areas of brain structures in prodromal Huntington's disease.

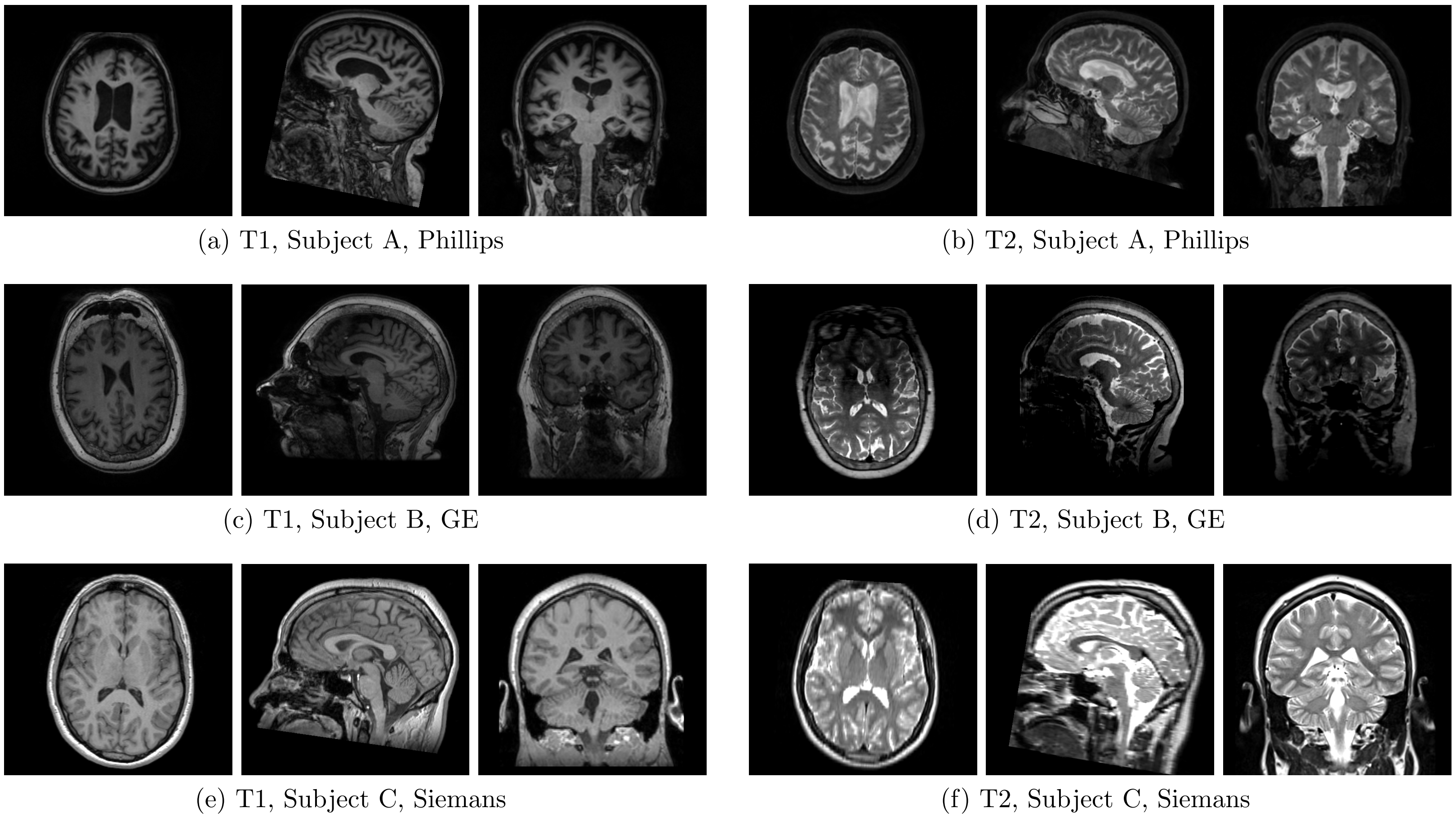
1.3 Challenges in developing a segmentation framework for large-scale and multi-center MRI dataset

> *With large datasets involving many variables, more structure can be discerned and variety of different approaches tried. What makes a dataset interesting is not only its size but also its complexity, and, perhaps most challenging, non-homogeneity; that is, different relationships hold between variables in different parts for the measurement space*
>
> *(* Leo Breiman, 1984 *)*

In order to have a successful automatic segmentation techniques, it is required to do careful planning, particularly to deal with heterogeneous data. While various algorithmic solutions to the MRI data analysis seem applicable, data inhomogeneity across scans confounds application of techniques on the large-scale data (See Figure 1.4). Based on our years of experience developing algorithms for multicentral data, various challenges encountered are described here, focussing on data inhomogeneity.

1.3.1 Inter-scan inhomogeneity: MRI artifacts

In MRI, artifacts are present for a variety of reasons and often causes intra-scan inhomogeneity, that make robust tool development difficult. Potential sources of artifacts include non-ideal hardware characteristics, intrinsic tissue properties and biological behavior, assumptions underlying the data acquisition and image reconstruction process, and poor choice of scanning parameters [147]. Careful study design



(a) T1, Subject A, Phillips

(b) T2, Subject A, Phillips

(c) T1, Subject B, GE

(d) T2, Subject B, GE

(e) T1, Subject C, Siemans

(f) T2, Subject C, Siemans

Figure 1.4: Example of Three Independent Scans from Predict-HD Study. Three individuals are scanned by using all different MR scanners. Considerable image inconsistency is observed with regard to 1) ventricle size differences (Subject A vs. B and C), 2) coverage of scan from head to neck, 3) inter-scan intensity inhomogeneity (by contrasting brightness between rows), 4) intra-scan intensity inhomogeneity (especially in Figure 1.4d), and 5) spatial orientation discrepancy.

and scanning protocols can attenuate the extent of artifacts from occurring, but some are unavoidable. A number of artifacts are listed in Table 1.4 paired with possible cause. MRI artifacts are one of the most concerning challenges in developing automatic MRI processing tools, that causes intra-scan inhomogeneity. That is, various kinds of MRI artifacts creates heterogeneous intensity profiles for identical tissue type depending on its spatial and relative location within the scanner. This differentiated MR intensities, intra-scanner inhomogeneous intensities pose a major hurdle for the robust tool development.

1.3.2 Intra-scan inhomogeneity: Data Variation by Multicenter and Longitudinal Study Design

Those multicenter data collection introduces a severe inter-scanner variation due to scanner calibration, software upgrades, field strength, procedural, and/or MR vendor differences. Therefore, even though multicenter a study provides several advantageous properties (See Section 1.2.4), processing and analyzing multicenteral data is often confounded by heterogeneous data across sites. Figure 1.4 shows three hardware examples from different sites, using independent MR sequence for three individuals. Distinction between three scans in the Figure 1.4 is very obvious even to the naked eye with respect to the brightness, scan coverage, intensity profiles, subject spatial orientation, and artifacts. These inter-scanner differences can be categorized as a non-random measurements error [61], which highly correlated to the site-specific characteristics and often degrades the data consistency further (Figure 1.4).

Table 1.4: MRI Artifacts and their cause

Artifact	Cause
RF Offset and Quadrature Ghost	Failure of the RF detection circuitry
RF Noise	Failure of the RF shielding
B_o Inhomogeneity	Metal object distorting the B_o field
Gradient	Failure in a magnetic field gradient
Susceptibility	Objects in the FOV with a higher or lower magnetic susceptibility
RF Inhomogeneity	Failure or normal operation of RF coil, and metal in the anatomy
Motion	Movement of the imaged object during the sequence
Flow	Movement of body fluids during the sequence
Chemical Shift	Large B_o and chemical shift difference between tissues
Partial Volume	Large voxel size
Wrap Around	Improperly chosen field of view
Gibbs Ringing	Small image matrix and sharp signal discontinuities in an image
Magic Angle	Angle between B_o and dipole axis in solids.

Various kinds of MRI artifacts exist and causing intra-scan intensity inhomogeneity with some degrees of variation. This table lists 13 common MRI artifacts and their cause from [69]. Heterogeneous intensity for same tissue type caused by artifacts is one of main challenges for MRI processing tool development.

Longitudinal study design brings in another issue with regard to data variation. As technologies advance, hardware are upgraded as well as software, and those upgrades lead to the MRI profile changes (Table 1.5). Even though those new technologies traded in for good, it often requires sophisticated development tuning to maintain the robustness of processing procedure. One of obvious dynamic occurring at MRI acquisition site is changes in manufacturer and field strength over several years of data collection period. Figure 1.5 shows distribution dynamic of collected data according to its MR manufacturer and field strength for PREDICT-HD study over years. The figure demonstrates MRI acquisition trend changes along time based on field strength and MR vendor and suggests a development process should take account those active changes as well.

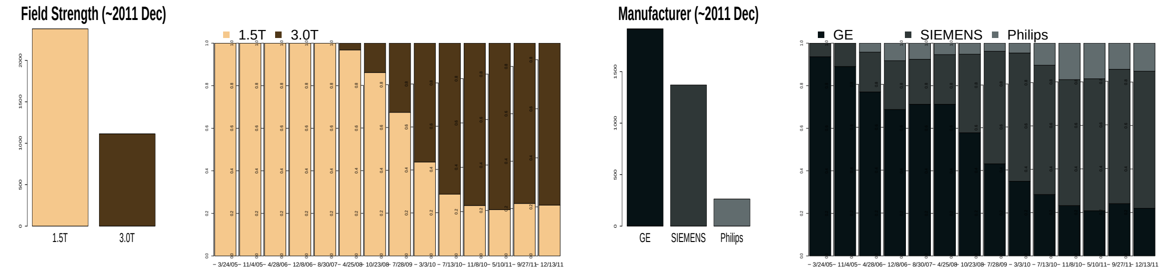
1.3.3 Large-Scale Data

In addition to the data variation, processing large-scale dataset becomes prohibitive with limited computer resources. The collected data in the PREDICT-HD project is over 3000 scans in 2013 May and the number is continuously growing. With our tool, average processing time for each scan session is roughly 3-5 hours depending on how many repeated scans collected at the session, in turn, $9,000$ to 15000 computer time requires.

In order to cope with all those limitations, we propose a segmentation pipeline method to overcome challenges to the large-scale, multi-center analysis study. The proposed method is clinically oriented to make use of sophisticated image-processing

Table 1.5: PREDICT-HD Study Source of Data Variation.

MR Vendor	Scanner Model	# Software Ver. Used
GE	Genesis signa	> 16
GE	Signa Excide	
GE	Signa HDX	
GE	Signa HDXt	
Philips	Achieva > 20	
Philips	Intera	
Siemens	Allegra	> 9
Siemens	Avanto	
Siemens	Espree	
Siemens	Sonata	
Siemens	Symphony	
Siemens	Symphony trim	
Siemens	TrioTrim	
Siemens	Verio	



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

Figure 1.5: This trend graph shows how MRI acquisition dynamic evolves with regard to 1) field strength and 2) MR vendor for PREDICT-HD study. The left two graphs are for field strength and the right two are for manufacturer of the scan. The 1.5a and 1.5c show the total portion of the collected scans between 2002 and 2011 for field strength and manufacturer respectively. The 1.5b and 1.5d represent the changes of the portion of the collected data as the study goes by from 2002. Even though the total proportion of data collected (1.5a and 1.5c) may infer more 1.5 Tesla data with the GE scanner, recent trend, which is sub-group of total data (in 1.5b and 1.5d), may differ from the entire data. That is, the recent trend from the graph demonstrates that the 1.5 Tesla data has been collected a lot less than before. In the same way, the manufacturer trend has been changed as well and note that the field strength evolves more rapidly than the manufacturer does. Those changes of MRI acquisition type leads to dynamical changes for the acquired scan properties, which, in turn, requires additional developmental effort.

methods and take the type and size of data into account. Figure 1.6 shows all the different types of scans that are collected under the PREDICT-HD study since 2002 across sites.

1.4 Metrics

Through out this study, manual traces of regions of interests are used as a gold standard for our development. **The quality segmentation results**, therefore, is measured as related to the gold standard, that can be estimated with multiple metrics , including relative overlap (RO), dice index (or dice similarity coefficient) (DSC), Hausdorff distance (HD), average Hausedorff distance (aHD) [41], intraclass correlation coefficient of absolute ($ICC(A)$) and consistency ($ICC(C)$).

- $RO = \dfrac{|A \cap B|}{|A \cup B|}$
- $DSC = \dfrac{2|A \cap B|}{|A| + |B|}$
- $HD = mD(A, B) = min_{a \in A}\{min_{b \in B}\{d(a, b)\}\}$
- $aHD = aD(A, B) = avg_{a \in A}\{avg_{b \in B}\{d(a, b)\}\}$

For ICCs, as a measure of the reliability between two different judgments, is employed to interpret our results in segmentation. Out of numerous definition of ICC,we used the $ICC(C, 1)$, which is consistency and $ICC(A, 1)$ that is absolute agreement followed the chart by McGRAW and Wong [97]. $ICC(C, 1)$ and $ICC(A, 1)$ can be corresponding to the definition of $ICC(3, 1)$, $ICC(2, 1)$ of Shrout and Fleissis [145] respectively

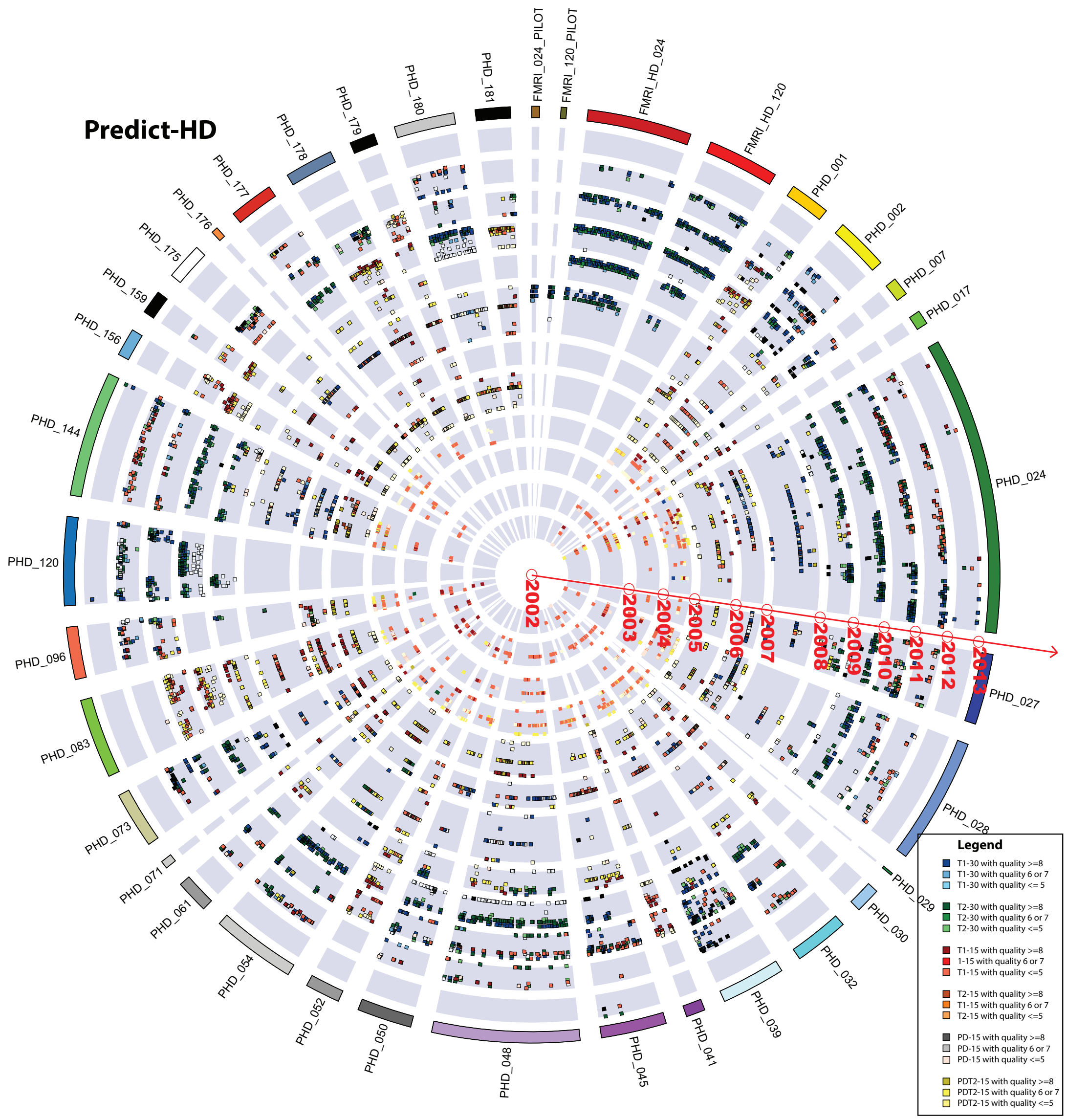


Figure 1.6: All the different types of scans that are collected under the PREDICT-HD study since 2002 across sites are represented.

1.5 Notation and Terminology

The description below describes notation and terminology followed in this report. Table A.1 also provides a quick reference of the notation and terminology.

1. **Image** $[\mathcal{I}]$. $\mathcal{I} : \mathbb{N}^2 \rightarrow \mathbb{R}$ for 2D and $\mathcal{I} : \mathbb{N}^3 \rightarrow \mathbb{R}$ for 3D.

2. **Feature Image Set** $[\mathcal{F}]$. A set of feature images that are given for input feature vector extraction: $\mathcal{F} \supset \mathcal{I}$. The most common feature image set would be T1- and T2-weighted images for the multi-modal MR data: $\mathcal{F} = \{\mathcal{I}_{T1}, \mathcal{I}_{T2}\}$.

3. **Voxel Location** $[i]$**.** A voxel location in the image $\mathcal{I}$ of size n. $i \in \{1, \cdots, n\}$.

4. **Feature Vector** $[\mathbf{f_i}]$**.** A set of representative descriptors at the voxel location i. Feature vector is also called as input feature vector, predictor variable, independent variable depending on project domains. $\mathbf{f_i} = (f_{i,1}, \cdots, f_{i,d})$

5. **Output Vector** $[y_i]$**.** A true output label that we would like to predict for the voxel i. For our brain MR segmentation problem, $y \in \mathcal{L}$ and $\mathcal{L} = \{l_1, \cdots, l_K\}$, where K number labels of interests. Output vector is also called as a response variable or dependent variable.

6. **Sample, Train, or Example data** $[\mathcal{S}_{n_s}]$**:** A given example data $s \in \mathcal{S}$ that machine-learning algorithm to be trained on. The training data s is an ordered pair of input feature $\mathbf{f}_i$ and the known output y_i at the image location i, where n is a number of total sample.

$$\mathcal{S} = \{s_i | s_i = (\mathbf{f}_i, y_i)\}$$

7. **Population Data**$[\mathcal{X}]$**.** Population data $x \in \mathcal{X}$ for a project of interest as a

super set of sample data $\mathcal{X} \supset \mathcal{S}$.

8. **Machine-Learning Model** [$\mathbb{M}$]**.** A machine-learning model predicting desired label based on a given feature vector. A trained machine-learning model, $\hat{\mathbb{M}}$, is a estimation of true model $\mathbb{M}$, which is specific to both algorithm and data.

1.6 Preliminary Work

This project is inspired from the former work by Powell et al. [121], that introduced an automated segmentation tool using artificial neural network. Recent years' increased interest in machine-learning for medical image processing have drawn our attention to investigate machine-learning based methods even further. Additionally, we reviewed the segmentation framework in general (Section 1.6.1), demonstrate the previous works [121, 84] according to specific enhancements to address multicenter problems (Section 1.6.2, 1.6.3, and 1.6.4) with the discussion about its limitations (Section 1.6.5).

1.6.1 Overview of the proposed Segmentation Framework

As shown in the Figure 1.7, the proposed segmentation framework consists of four distinct phases with two pre-processing stages. Two pre-processing stages, *spatial normalization* [54] and *bias field correction* [83], facilitate more reliable operations of the entire framework by providing consistency of the MR data, and they proved their benefits for the reliable analysis of multi-site MR data in [119]. Briefly, the *spatial normalization* rigidly aligns each MR data so that anterior-commisure (AC) and posterior-commisure (PC) are placed at the main axis with the AC point as an origin.

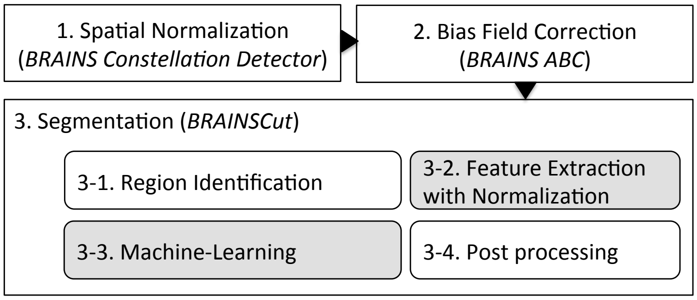

1. Spatial Normalization
(BRAINS Constellation Detector)
2. Bias Field Correction
(BRAINS ABC)
3. Segmentation (BRAINSCut)
3-1. Region Identification
3-2. Feature Extraction with Normalization
3-3. Machine-Learning
3-4. Post processing

Figure 1.7: Segmentation Framework Overview: Main segmentation process, *BRAINSCut*, employs the results of two preprocessing steps: *spatial normalization* and *bias field correction*. The main segmentation process then comprises of four stages of *3-1. region identification*, *3-2. Feature extraction with normalization*, *3-3. Machine-learning*, and *3-4. post processing*. Original segmentation framework, where this presented work is stemmed from, are explained in [122, 84]. In this paper, we aim to determine the robust components of segmentation framework for large-scale multicenter data analysis.

For the robustness of subsequent processing, this spatial normalization minimizes inconsistency of MRI's anatomical orientation across scans. *Bias-field correction* then maximizes intra-scan intensity homogeneity by taking advantages of multiple scans (multi-modal if applicable) collected at each site.

The core segmentation framework then begins with *candidate region identification* [122, 84]. *Template spatial priors* are created by warping and averaging a reference set of the manual traces into the template atlas [157]. The candidate region is then identified by *subject-specific priors*, which are generated by deforming the template (region-specific) spatial priors to the subject-specific space. The warping used in this work has been significantly improved over our previous work by implementing the robust high-deformable registration from the Advanced Normalization Toolkit (ANTS) [6].

Next, the *feature extraction* stage samples only voxels that have a spatial probability in the range of $(0, 100)$ percent [122, 84] based on the *subject-specific spatial prior*. Elimination of both $< 1\%$ and $> 99\%$ probability regions allows controlled ML training only for uncertain voxels. A ML algorithm then classifies each of the uncertain voxel to a specific label, and finally *post-processing* completes the segmentation framework. The post-processing is used to address undesirable small anomalies that may occur due to inherent noise in the MR data. For example, small interior holes in the ML defined structure are filled using a standard morphological hole filling algorithm. These four stages of proposed segmentation framework were carefully designed and validated specifically for scalable multi-site longitudinal data processing. In the following sections, we describes each enhancement component in more detail.

1.6.2 Robust Candidate Region Identification

One of key features in our segmentation framework is a candidate region identification step (See Figure 1.7). For brain MR image, instead of searching whole image for our relatively smaller regions of interests, search area is localized by placing region-specific spatial priors defined in a common atlas space on top of subject-specific space. The localization of search area accelerates executing time and reduces false positives by effectively restricting search area. The candidate region identification process also allow machine-learning algorithm to concentrate on those relevant voxels that effectively characterizing our regions of interests. Example of region-specific priors for caudate and putamen in left hemisphere are shown in Figure 1.8. How robustness of high-deformable registration was enhanced for the candidate region identification process (Section 1.6.3) are described in the following sections.

Region-specific spatial priors generated by averaging all the manual traces of the training set on the template space and then warped to the subject space by utilizing registration between $T1$-weighted images of subject and template [84]. Spatial priors have values between zero and one, respectively meaning 0 and 100% possibility being the structure. A Gaussian smoothing operation was applied on a set of averaged manual traces to take account of anatomical variation in larger unseen MR data. Again, this localization encourages robust processing of subsequence steps by retrieving only relevant information (See Figure 1.8), in addition to reducing training size for effectiveness.

1.6.3 Registration reinforcement with Robust Initialization

Accurate candidate region identification depends on high quality registration. Of multiple available choices, high-deformable registration produced the most accurate segmentation results for our framework. Finding a robust and accurate high-deformable registration method, however, is an open and complicated research problem. Especially for large-scale MR data with substantial variation, optimizing parameters for a registration algorithm is very demanding issue. One complication we often encountered is that a set of parameters that works for one sites image intensity profile ften fails on another due to unexpected field of view or profile variation. In other words, it is difficult issue to find parameter set that *globally* works for large-scale

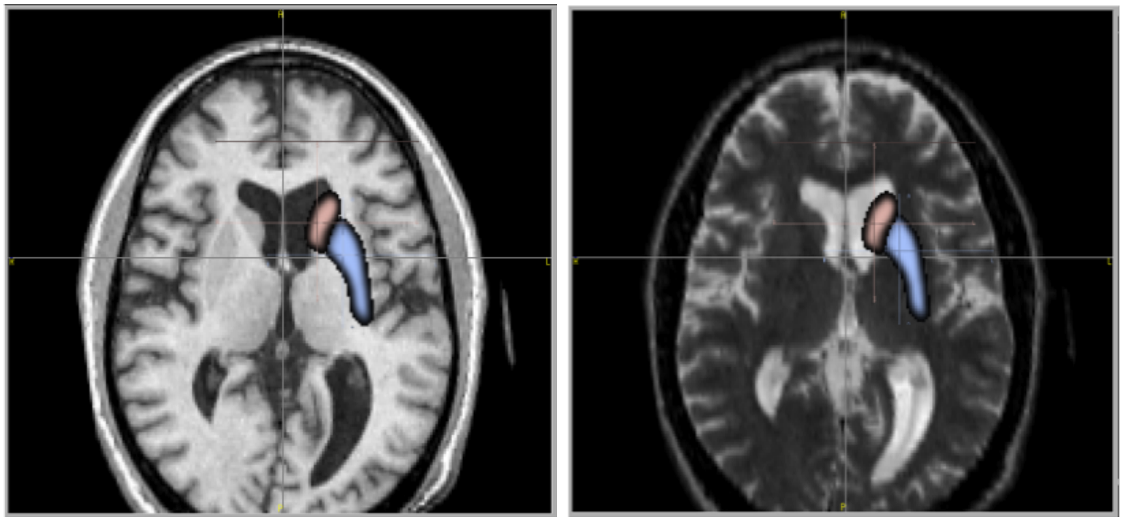


Figure 1.8: **Region-specific spatial priors** are generated from 32 manual traces for six subcortical structures to localize search regions of interest and the figure shows example priors of caudate (red) and putamen (blue) in left hemisphere overlaid on top of template $T1$ (left) and co-aligned $T2$ (right) MR images. From 32 manual traces, each of labels is warped into template $T1$ space and then averaged over to create template *region-specific spatial priors* Smoothing operation were performed on each averaged priors via Gaussian filter to encompass greater anatomical variability in larger and unseen population data.

data sets.

Landmark-based initialization was devised for the high-deformable registration to ensure robustness of warping and to minimize parameter searching effort, especially for large-scale data with substantial data variation. In general, high-deformable (non-rigid) registration excels in computing correspondence between inter-subject morphological differences than lower level warping methodologies, such as rigid and affine transformation. The high-deformable transformation algorithm, however, often converges to a locally optimal solution if the given initial condition is poor. That is, it is often true that non-rigid methods require a sufficiently good initialization to converge on a good final solution. The initialization with landmarks is employed to avoid those undesired convergence at local minima as well as to increase efficiency by providing a sufficiently good initial solution and consistent field of view.

Landmarks detected by *BRAINS Constellation Detector* module [54] (See Figure 1.9) are utilized to provide a good starting point for the high-deformable registration. The algorithm estimating affine transform as described in [150] that are implemented and integrated by the ITK community [82], to get the robust initialization from landmarks computed for $T1$-weighted images. This landmark based initialization process enhances robustness of registration algorithm by providing a good initial solution and also increase convergence speed as well.

The work in [150] describes mathematics for calculating affine transformation from the number of paired landmarks. The algorithm will calculate the best affine transform based on the least squares sense. For the two sets of points P and Q given

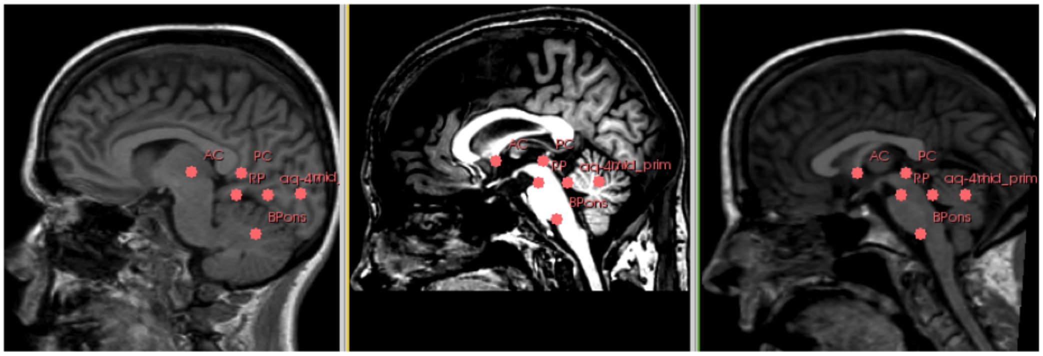


Figure 1.9: Landmarks detected on each subject have been utilized to estimate the affine transformation between template and subject $T1$-weighted images. The estimated affine transformation is employed as an initial transform for high-deformable registration to increase robustness of the algorithm for large-scale multicenter data processing.

by

$$p_i = (p_{1i}, ..., p_{ni})^T, q_i = (q_{1i}, ..., q_{ni})^T (i = 1, ..., m)$$

The author derives equation to get two matrices, A and t, to satisfy

$$p_i \approx Aq_i + t(i = 1, ..., m)$$

After some calculation and rearrangement, we get the following form of equation:

$$\tilde{Q}\tilde{a}_j = \tilde{c}_j(j = 1, ..., n)$$

where

$$\tilde{q}_i = (q_{1i}, ..., q_{ni}, q_{n+1,i})^T, \text{where } q_{n+1,i} = 1$$

$$\tilde{Q} = \sum_{i=1}^{m} (\tilde{q}_i \cdot \tilde{q}_i^T)$$

$$\tilde{a}_j = (a_{j1}, ..., a_{jn}, t_j)^T$$

$$\tilde{c}_j = (\tilde{c}_{j1}, ..., \tilde{c}_{j,n+1})^T \text{with } \tilde{c}_{jk} = \sum_{i=1}^{m} (q_{ki} p_{ji})(k = 1, ..., n + 1).$$

For more mathematical details, please see [150].

Weighting is simply added to the calculation by substituting

$$p_i' = W \cdot p_i$$

$$q_i' = W \cdot q_i,$$

where W is $n \times n$ diagonal matrix with weights in it. Figure 1.10 and Figure 1.11 describe toy examples that estimating affine transformation by using the ITK implemented function [82]. Toy example was first tested with five paired points ((a)-(e)) for two conditions: 1) even weights and 2) adjusted (more) weights on (a) and (e), and results came out as expected. That is, for (a) and (e), they showed better correspondence when they are heavily weighted than others (left graph of Figure 1.10). The second toy example was designed to see the effect of noisy and/or uncertain landmarks (symbolized as (f) and (g) in Figure 1.11) in the affine estimation. As expected, there was no or minimum effect was observed from those noisy points when their weights were set to very small. Note that all the weights are relative each other.

1.6.4 Robust Feature Extraction

For each voxel i, feature vectors are extracted for locational information in symmetry spherical coordinate information [122] (Section 1.6.4.1), for intensities in

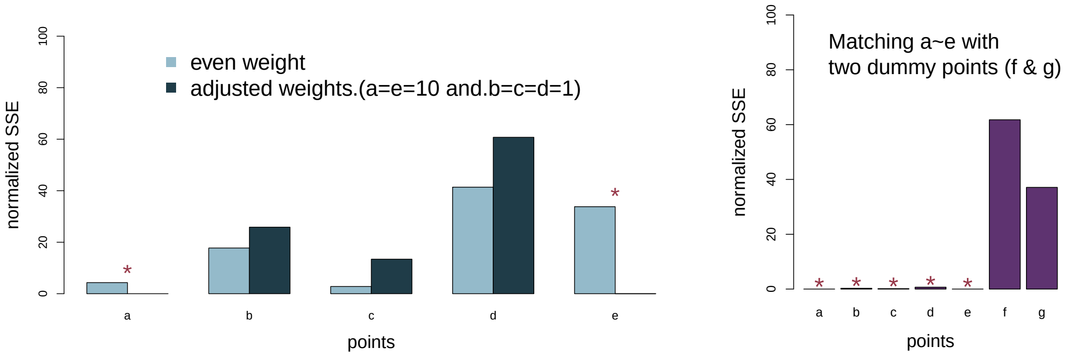


Figure 1.10: Algorithm estimating affine transformation was implemented to improve initialization of high-deformable registration algorithm. A pair of landmarks are provided as a toy example to demonstration how *weighting* behaves on the estimation. Affine transform from moving to fixed points are estimated and compared between three conditions: c1) default even weights for all five points (graph on left-hand side), c2) adjusted weights- more weights on (a) and (e) (graph on left-hand side), and c3) with dummy points of (f) and (e) (graph on right-hand side), which simulate noisy and/or uncertain landmarks and thus less important landmarks than other five in the estimation. Appearance of each moving and fixed points are displayed in Figure 1.11. This toy example demonstrates that affine transform can be estimated with minimal effect from noisy data by weighting scheme.

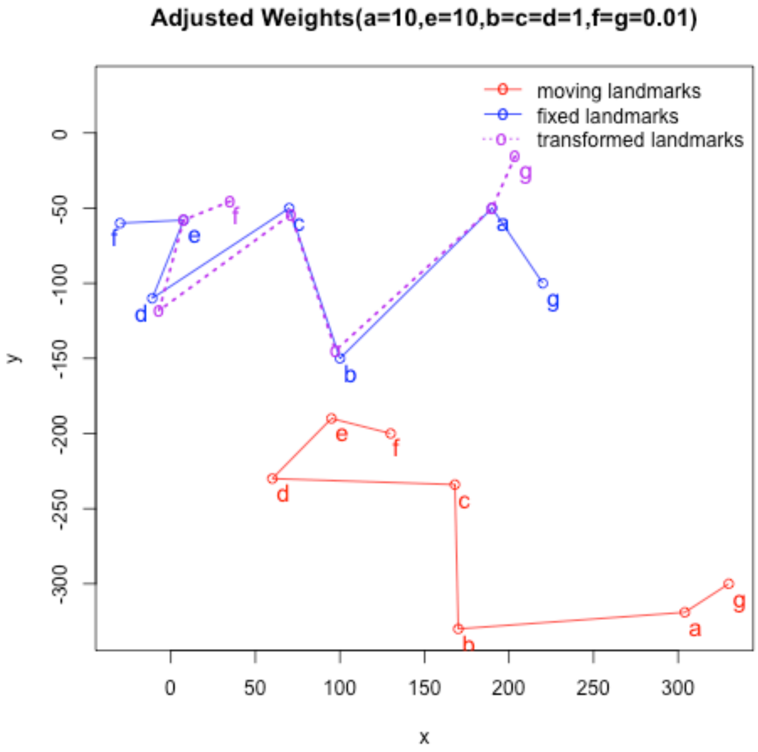


Figure 1.11: This figure illustrates how noisy landmarks can be under-weighted so that their effect on estimation of affine transformation can be minimized. The figures shows estimated affine transformation from moving (blue) points to the corresponding fixed (red) points. Warped points (purple) are well aligned to the corresponding points (blue) except for points (f) and (g), which simulating noisy, or uncertain landmarks. As shown in Figure 1.10, normalized sum of squared error (SSE) was a lot smaller for the points of interest, (a) (e), than the dummy points (f) and (g), which simulating noise in vivo data. This figure supports our hypothesis that less weighted landmarks are aligned less perfectly while preserving correspondence of points of interests ((a)-(e)).

neighbors along the gradient descent along deformed priors (Section 1.6.4.2), and for the candidate vector based on deformed region-specific priors (Section 1.6.4.3).

1.6.4.1 Symmetrical Spherical Coordinate Information

Spherical coordinate definition is employed to provide locational information relative to AC point for the voxel i:

$$\mathbb{S}_i = \{\rho_i, \phi_i, \theta_i\} \tag{1.1}$$

Note that, in the previous study [84], the definition of traditional spherical coordinate system was modified to take account of brain's symmetrical morphology between left and right hemisphere (See Figure 1.12 and Figure 1.13). The modified version of symmetrical spherical coordinates definition is:

$$\rho = \sqrt{x^2 + y^2 + z^2} \tag{1.2}$$

$$\phi = arctan(\frac{|x|}{y}) \tag{1.3}$$

$$\theta = arctan(\frac{|z|}{y}). \tag{1.4}$$

The modification of spherical coordinate information is motivated by multiple failure cases (Figure 1.14) that we observed from our experiments. By propagating information about biological symmetry in human brain along to the AC-PC line into the learning process, the failure cases could be saved and results are contrasted in Figure 1.14 between original and modified definition of spherical coordinate system.

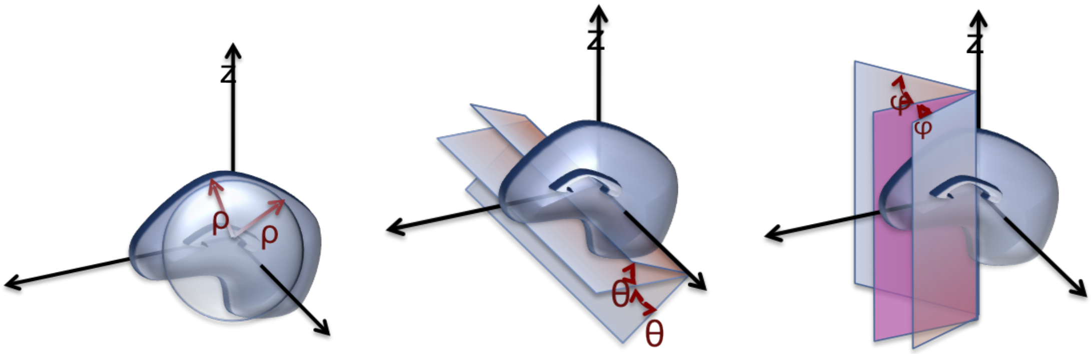
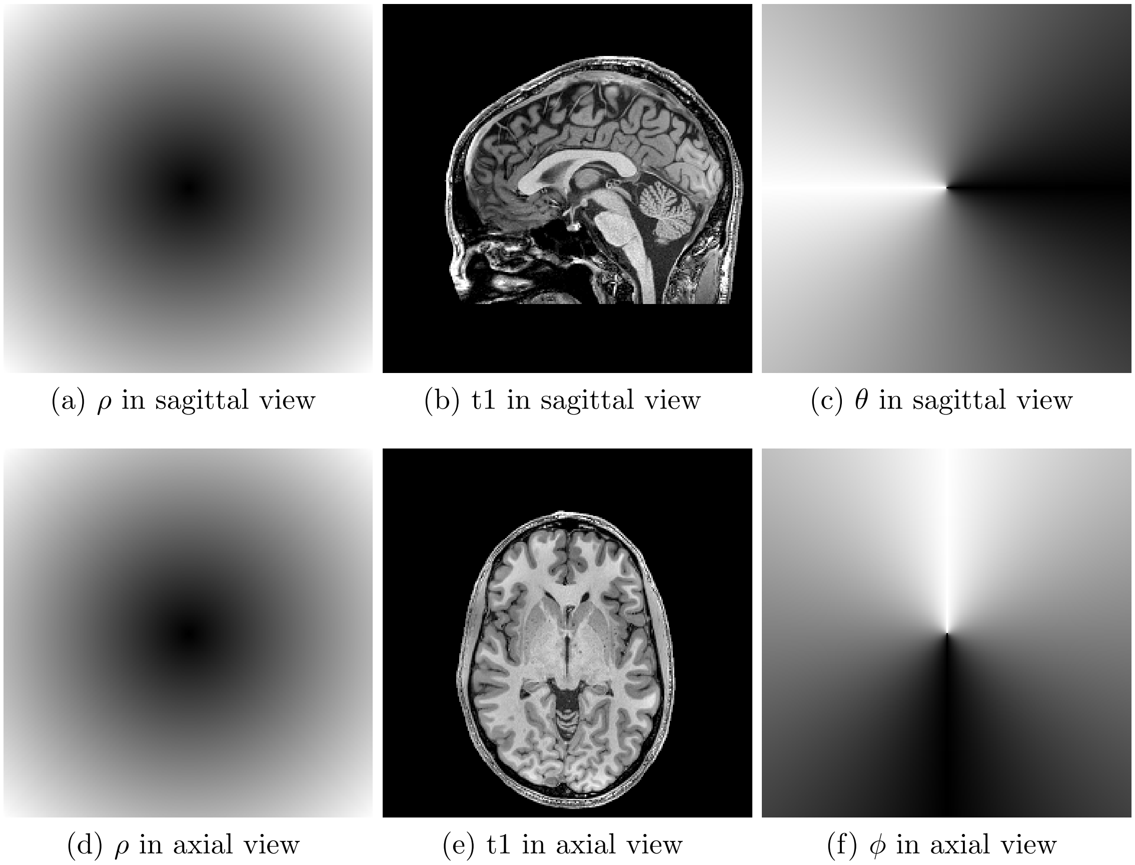


Figure 1.12: Symmetrical spherical coordinate definition is illustrated to take account of brain symmetry between left and right structures. A new symmetry definition of each ρ, ϕ, and θ is shown in the Equation 1.2-1.4. We have identified failure cases due to the discontinuity and asymmetry of those classical definitions and modification improved segmentation results.

1.6.4.2 Neighbors along the Gradient Descent of Deformed Region-Specific Priors

Intensity values are extracted for each voxel at i for all the images given $I_i \in \mathcal{I}$ including neighbors selected along gradient descent direction of deformed priors. Instead of using classical definition of 6-neighbor model (See Figure 1.15a), directionally consistent neighbors, that sampled along the gradient descent direction of deformed region-specific priors are employed for preserving biological directional consistency *from inside to outside* of structures (See Figure 1.15b). This direction can also be thought as a normal of the surface of structures so that the machine-learning model can be trained for right surface. The input feature vector G along the gradient


(a) ρ in sagittal view

(b) t1 in sagittal view

(c) θ in sagittal view

(d) ρ in axial view

(e) t1 in axial view

(f) ϕ in axial view

Figure 1.13: Computed symmetrical spherical coordinate at the template $T1$ space, regarding AC-PC as y, mid-sagittal plane as z, and the orthogonal plane as x axis. Each ρ, ϕ, and θ image is displayed with template $T1$ image in the same row. A new symmetry definition of each ρ, ϕ, and θ is shown in the Equation 1.2-1.4. We have identified failure cases due to the discontinuity and asymmetry of those classical definitions and modification improved segmentation results.

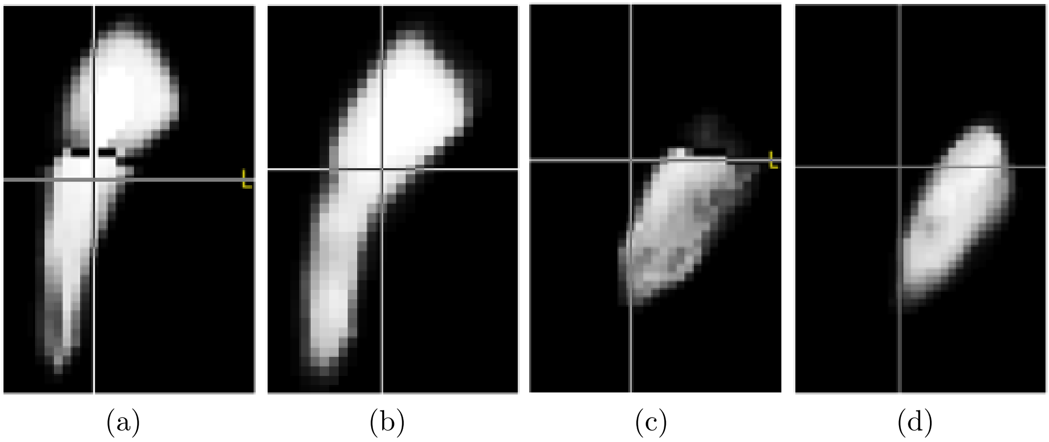


Figure 1.14: Symmetrical definition of spherical coordination input was devised to take account of brain symmetry in left and right hemisphere. Failure cases of putamen (1.14a and globus pallidum (1.14c) improved with our new symmetry definition for both putamen (1.14b) and globus pallidum (1.14d).

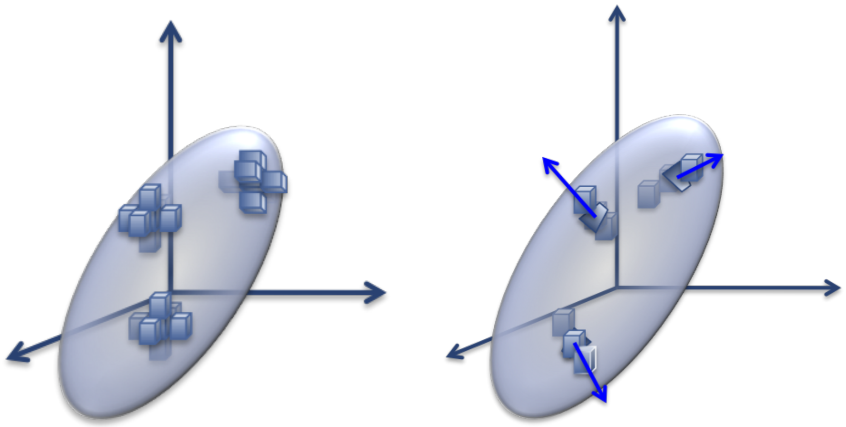
descent direction can be written:

$$G = G_{i,\mathcal{I}_j}, \text{ where } \forall \mathcal{I}_j \in \mathcal{F} \tag{1.5}$$

$$G_{i,\mathcal{I}_j} = \{\mathcal{I}_j(i - G1),\ \mathcal{I}_j(i),\ \mathcal{I}_j(i + G1)\}, \tag{1.6}$$

where $i - G1$ and $i + G1$ is the voxel location that are located at one voxel inward and outward along the gradient descent from the location i.

Additional advantage of this sampling approach is the *efficiency*. The neighborhood definition along the deformed region-specific gradient descents only creates one directional neighbors, that reduces the neighbor size by factor of three, comparing to six-neighborhood scheme. The reduced number of input vector size saves time required for training as well.



(a) Classical Neighbor (b) Locationally Invariant Neighbor

Figure 1.15: In this study, neighborhood information of each voxel is extracted along the gradient descent direction of deformed region-specific priors for all the input images given. Classical definition of neighborhood of the voxel (Figure 1.15a) ignores relative locational knowledge between neighbors. Extracting neighbors in input feature vectors along gradient descent direction of priors (Figure 1.15b), which results in sampling neighbors from inside to outside direction, ensures consistent feature sampling, that provides consistent relative locational information. Further more, the neighborhood along the gradient descent direction reduces the number of element in the vector by factor of three, which improves efficiency of algorithm.

1.6.4.3 Feature of Candidate Vector

The last component of input feature is *candidate vector* that identified from deformed spatial priors. Based on warped priors on top of a subject space, candidate vector was expressed in Boolean vector. As an ordered Boolean vector, each element expresses if the voxel is possible to be the structure. If a prior value of specific structure at location i has value bigger than zero, the Boolean value is set to true, otherwise false. Therefore, candidate feature vector $\mathcal{C}$ at a voxel location i can be written as following:

$$\mathcal{C}_i = \{\mathbb{1}_{l_1}, \cdots, \mathbb{1}_{l_K}\}$$

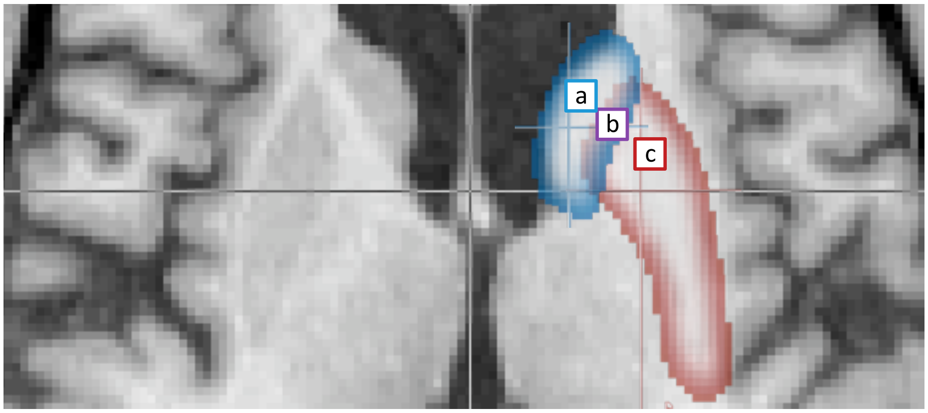


Figure 1.16: Candidate vector computation example based on two deformed region-specific priors of left caudate (blue) and left putamen (red) in the boolean format. This feature is especially useful in multi-structural training since this feature provides information where the voxel could belong to. For three voxels shown in this figure, a, b, and c, candidate vector C at voxel i $\mathcal{C}_i = \{\mathbb{1}_{\text{left caudate}}, \mathbb{1}_{\text{left putamen}}\}$ of each voxel is $\mathcal{C}_a = \{True, False\}$, $\mathcal{C}_b = \{True, True\}$, and $\mathcal{C}_c = \{False, True\}$, respectively.

In summary, a feature vector $\mathbf{f_i}$ for voxel i is given

$$\mathbf{f_i} = \{\mathcal{S}_i,\ G_{i,\mathcal{I}},\ \mathcal{C}_i\}, \tag{1.7}$$

where $\mathcal{S}$ is a symmetrical spherical coordinate information (Equation 1.1), $G_{i,\mathcal{I}}$ is image intensity along the gradient descent direction of deformed prior at the image location i (Equation 1.5) [122, 84] for $\mathcal{I}_j \in \mathcal{F}$. Note that our previous work usually utilized a feature image set $\mathcal{F} = \{\mathcal{I}_{T1}, \mathcal{I}_{T2}, \mathcal{I}_{SG}\}$ when SG is a sum of gradient magnitude image of $T1$ and $T2$ weighted MRI scans. The samples were created independently for each of the subcortical structures in both left and right hemisphere: $\mathcal{C}_i = \{\mathbb{1}_{left\ ROI}, \mathbb{1}_{right\ ROI}\}$

1.6.5 Results and Discussion

We revisit the results of the previous study [84] in Section 1.6.5.1 and further elaborate the related discussions in Section 1.6.5.2 with updated literature reviews. The expanded discussion is to legitimate the direction of this development from this report by providing in-depth understanding of success and challenges of our previous reports. We conclude this section with summary of this chapter in Section 1.6.5.3.

1.6.5.1 Improvements on Single-Site Data with Enhancements

24 data sets of multi-modal, $T1$ and $T2$, MR images were selected from our project of interest domain at a *single* site, blinded to any clinical information, such as disease progression, age, and gender. Each scan, however, was chosen to have various characteristics based on non-clinical knowledge, such as brain volume/size, and rough

ratio of tissues so that our training set thoroughly spans the anatomical variations encountered in our study domain, and was not biased towards one particular configuration. *Hold-out* test were performed to assess performance of segmentation. Out of 24 data sets, 16 data sets were chosen randomly for training and 8 data sets were reserved for testing purposes.

The comparison result was provided only for four structures of caudate, hippocampus, thalamus, and putamen due to the limited availability of results in the previous work [122]. Note that the work presented in [84] provides a complete set of study for all six subcortical structures. Performance of all six subcortical structures are reported in Table 1.6 independently, which displays high segmentation correspondence to the manual traces in general.

In the study [84], we concluded that utilization of feature-enhanced images, which include a soft-tissue classified image and mean of gradient magnitude image, improves reliability of our segmentation. It was also quantitatively shown that the volumetric overlaps with manual traces has been increased from the former work [122]. The result was reliable compared to the manual segmentation. Structures with a relatively small volumes (nucleus accumben and hippocampus) or vague intensity boundaries (globus pallidum, thalamus and also nucleus accumben) were successfully identified for single-site data even with relatively small set of training data.

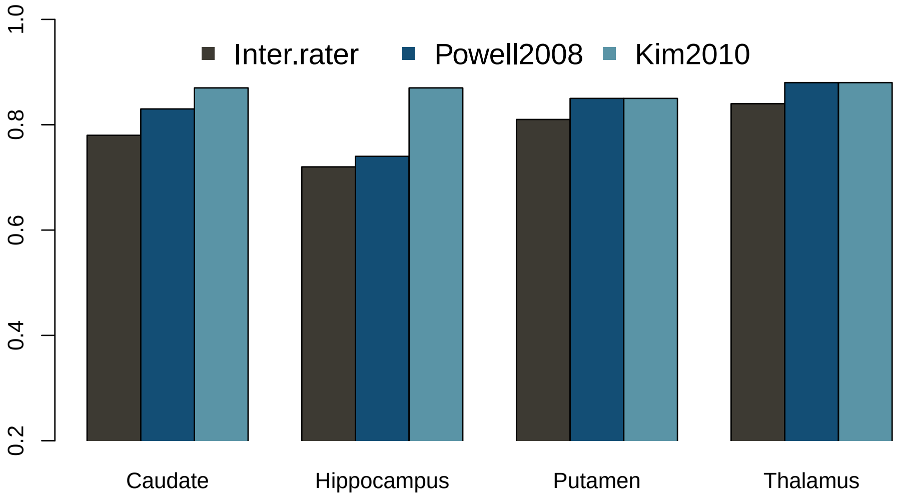


Figure 1.17: Segmentation accuracy in terms of relative overlap (RO) to the manual traces are advanced with three incorporated enhancements to the input feature vector as described in [84] and Section 1.6.4 from the work in [122]. Also note that ROs to manual traces from both automated segmentation results by Powell et al. [122] ($Powell2008$) and Kim et al. [84] ($Kim2010$) are well above to the inter-rater correspondence. This well supports that incorporated enhancements ($Kim2010$) including 1) symmetrical spherical coordinate, 2) candidate vector , and 3) additional feature image SG advanced segmentation accuracy.

Table 1.6: Segmentation results

ROI	$\overline{\lvert A\rvert - \lvert M\rvert}$	*Std*	*RO*	*DSC*	*ICC(A)*	*ICC(C)*	*Pearson r*
accu	337.75(4.9%)	60.841(19.6%)	0.750 ± 0.091	0.734 ± 0.078	0.73	0.75	0.76
caud	3294.6(3.7%)	419.05(29.7%)	0.878 ± 0.041	0.879 ± 0.039	0.78	0.8	0.83
glob	1573.3(-0.5%)	248.72(26.9%)	0.799 ± 0.063	0.794 ± 0.060	0.8	0.79	0.81
hipp	2188.3(-0.5%)	259.58(5.5%)	0.870 ± 0.047	0.838 ± 0.050	0.91	0.91	0.91
puta	4796.6(2.7%)	472.67(45.2%)	0.850 ± 0.046	0.857 ± 0.046	0.73	0.75	0.81
thal	6495.9(1.1%)	809.63(6.8%)	0.877 ± 0.038	0.876 ± 0.045	0.89	0.89	0.89

Segmentation results by using the method described in [84] from single-center data set showing high correspondence to manual traces in regarding all seven metrics for all six subcortical structures. The table reports volume differences ($\overline{|A| - |M|}$), standard deviation (Std), relative overlap (RO), dice index (DSC), intraclass correlation of absolute ($ICC(A)$) and consistency ($ICC(C)$). Especially both $ICC(A)$ and $ICC(C)$ were above 0.75, that commonly regarded as a rule of thumb suggested in [145] for both raters' delineation to be considered as identical.

1.6.5.2 Issues of Multi-structural Segmentation with regard to Imbalanced Data

In the previous study [84], we also investigated segmentation performance of simultaneously constructed models for all six subcortical structures only briefly. Even though the improvements that we devised partially succeeded, the results (comparing the single-structural result in Table 1.6 to multi-structural one in Table 1.7: See Figure 1.18) were **inconclusive** for which approach is superior. In this regards, we expand the discussion about multi-structural, or multi-label, classification issue for our segmentation framework in conjunction with those ambiguous results.

In the previous study [84], with some restrictions, we concluded that the utilization of multi-structural model construction from ANN is effective in that reduces time requirement for the individual model building of six ROIs. In contrast to our expectation, however, the simultaneously trained model were not significantly superior to those single structure models (Figure 1.18). We interpreted that the inferior result possibly due to complicated parameter adjustment for multiple structures in one model construction, such as number of hidden nodes and threshold value.

Observing those performance drops with increased number of ROIs for the multi-structural model construction [84] and also further literature review lead us to a *class imbalance* [100, 73]. As noted in [73], the class imbalance problem is a relative problem that depends on 1) the degree of class imbalance, which increases as number of ROIs increases; 2) the complexity of the concept represented by the data; 3) the

Table 1.7: Multistructural segmentation results by using ANN

ROI	$\overline{\|A\| - \|M\|}$	*Std*	*RO*	*DSC*	*ICC(A)*	*ICC(C)*	*Pearson r*
accu	267.56(-16.9%)	59.431(16.8%)	0.724 ± 0.097	0.690 ± 0.100	0.59	0.87	0.88
caud	3181.7(0.1%)	371.08(14.8%)	0.868 ± 0.043	0.872 ± 0.042	0.80	0.80	0.80
glob	1409.3(-10.9%)	179.89(-8.2%)	0.757 ± 0.084	0.779 ± 0.069	0.47	0.65	0.66
hipp	2136.1(-2.8%)	270.91(10.1%)	0.860 ± 0.050	0.829 ± 0.053	0.89	0.91	0.91
puta	4688.8(0.4%)	479.48(47.3%)	0.839 ± 0.050	0.848 ± 0.047	0.70	0.69	0.70
thal	6442.5(0.2%)	781.93(3.2%)	0.880 ± 0.039	0.882 ± 0.044	0.90	0.90	0.90

Multistructural segmentation results by using ANN [84] in terms of mean volume differences ($\overline{|A| - |M|}$), standard deviation (Std), relative overlap (RO), dice index (DSC), intraclass correlation of absolute ($ICC(A)$) and consistency ($ICC(C)$), and pearson's r. A comparative graph is given in the Figure 1.18, where performance preference is not conclusive.

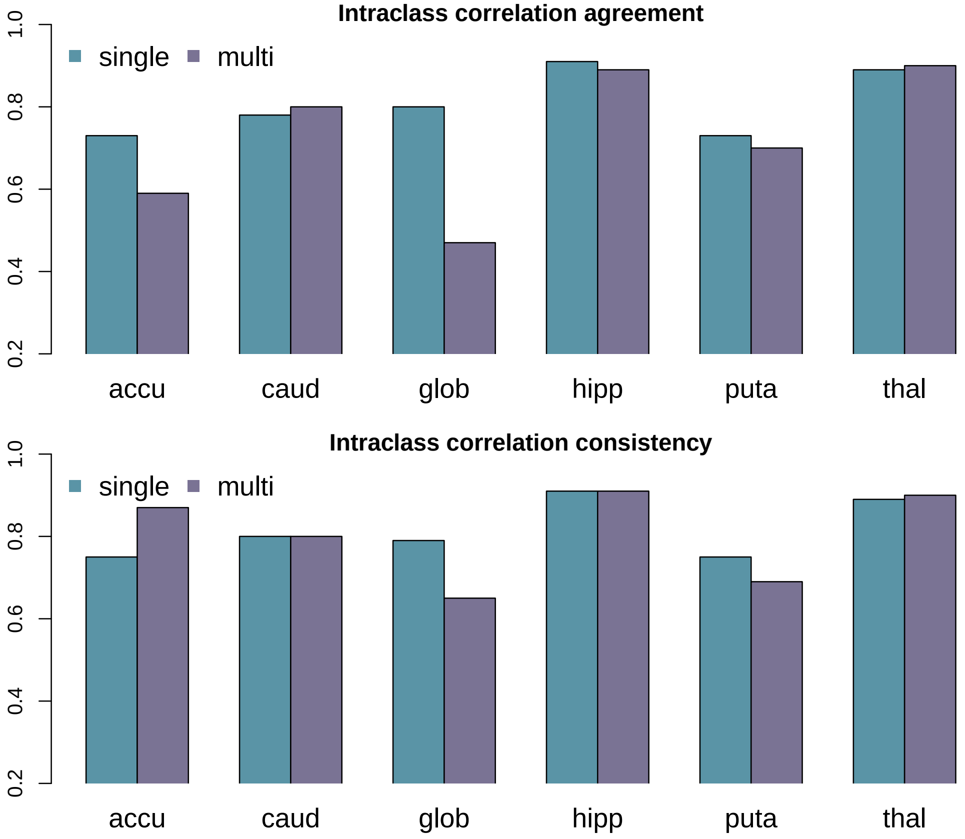


Figure 1.18: This multi-structural segmentation results produced by the method described in [84] is shown in comparison to the sigle structural trial in terms of intraclass correlation of absolute (top) and consistency (bottom) for nucleus accumben (*accu*), caudate (*caud*), globus pallidum (*glob*), hippocampus (*hipp*), putamen (*puta*), and thalamus (*thal*). In the study, we summarized that the multi-structural machine-learning model may provide an effective approach to minimize development time for each individual structures. Further literature review and experience about multi-structure classification with machine-learning, however, sugggests that there is a serious limitation for the method to obtain the identical degree of performance, class imbalanced issue. It is also known that imbalanced class issue can be more crucial for the intricate problem.

overall size of the training set; and 4) the classifier involved. The study in [73] also reported that the higher the degree of class imbalance the higher the complexity of the concept and the smaller the overall size of the training set, the greater the effect of class imbalances in classifiers sensitive to the problem.

In our cases, we diagnosed that our class imbalance issue may come from two factors: 1) size differences among ROIs, and 2) region and non-region ratio by candidate region identification. As the number of ROIs increases, those two factors join together and accelerates asymmetry among structures. The sophisticated concept of each brain structures also makes it harder for multi-label classifer to produce compatible segmentation accuracy to the single-label classifier.

Furthermore, it is also reported that misclassify examples of the minority class is more costly than misclassify examples of the majority class [100]. That is, smaller structures, and thus often intricate to identify, are more vulnerable to the misclassification presented in the training data for constructing the multi-structural machine-learning model. In those respects, we do not introduce any further investigation on multi-structural segmentation approaches in this report. The studies introduced in this report only focused on uni-structural segmentation methodologies of using machine-learning classifiers.

1.6.5.3 Summary and Conclusion

We reviewed related background and three preliminary enhancements for our subcortical segmentation framework: **1)** incorporation of the robust candidate region identification step (Section 1.6.1) **2)** incorporation of the high-deformable registration with the accurate landmark initialization (Section 1.6.3), **3)** incorporation of the robust feature extraction using symmetrical spherical coordinate definition, directionally consistent neighbor information, and introduction of candidate vectors (Section 1.6.4). Note that the neighborhood sampling along the gradient descent of deformed spatial priors (Section 1.6.4.2) already integrated and validated its effectiveness in the study [122]. The later study results [84] also well supported advantageous effects of those enhancements for the segmentation pipeline by comparing to the former one [122] (See Figure 1.17). The multi-structural segmentation was also tried to see the effect of simultaneous segmentation of neighboring structures [84]. The segmentation accuracy was evaluated against manual traces to ensure the clinical validity of acquired volumes.

CHAPTER 2
ROBUST MULTI-SITE MR DATA PROCESSING FOR AUTOMATED SEGMENTATION FRAMEWORK: ITERATIVE OPTIMIZATION OF BIAS CORRECTION, TISSUE CLASSIFICATION, AND REGISTRATION

A robust pre-processing multi-modal tool, for automated registration, bias correction, and tissue classification, has been implemented for large-scale heterogeneous multi-site longitudinal MR data analysis. This work focused on improving the iterative optimization framework between bias-correction, registration, and tissue classification inspired from work of others [172, 166, 8]. Our hypothesis is that robust bias correction will result in improved segmentation results by providing intensity-consistent MR data across subject and sites. The primary contributions are robustness improvements from incorporation of following four elements: 1) utilization of multi-modal and repeated scans simultaneously, 2) incorporation of high-deformable registration, 3) use of extended set of tissue definitions, and 4) use of multi-modal aware intensity-context priors. The benefits of these enhancements were investigated by a series of experiments with both simulated brain data set (BrainWeb) and by applying to highly-heterogeneous data from a 32 site imaging study with quality assessments through the expert visual inspection. In addition, we applied our subcortical segmentation framework to contrast between the before and after of bias correction. The segmentation comparison suggested that bias corrected input improves segmentation accuracy in general. The implementation of this tool is tailored for, but not limited to, large-scale data processing with great data variation with a flexible interface. With these enhancements, the bias-corrected images showed improved robustness for large-scale heterogeneous MRI processing and segmentation accuracy is enhanced for six subcortical structures.

2.1 Introduction

A key research technique for advancing the understanding of the human brain is the analysis of large collections of MR images [1]. Accurate and robust analysis of brain MR imaging from multi-site, multi-modal and longitudinal studies is a difficult problem, significantly confounded by intensity inhomogeneity across site, modality, and time. Variations in intensity due to data collection from different scanner manufacturers and scanning environments are a primary challenge associated with automating the analysis of those studies. The development of techniques to address large variabilities in scan properties, such as high-quality registration and bias-field correction become essential to ensure interpretation consistency among these datasets. As more research begins to use this data collection model, there has been an increased emphasis on automated tool development that addresses the challenges of multicenter MR image analysis. Successful development of a fully automated analysis framework can reduce both the operator time requirement and measurement variability for clinical trial applications [182].

Iterative approaches are attractive because they naturally become one inter-

[1] Examples of data: ADNI database `http://www.adni-info.org`, OASIS `http://www.oasis-brains.org`, IXI dataset `http://www.braindevelopment.org`, INDI `http://fcon\_1000.projects.nitrc.org/fcpClassic/FcpTable.html`, and BrainWeb `http://www.bic.mni.mcgill.ca/brainweb/`

related and interconnected optimization problem. Iterative optimization approach is proposed to achieve robust MR processing, often involving three main techniques: bias-field correction, registration, and tissue classification. The improved intensity uniformity provided by bias-field correction produces better registration accuracy, and also enhances tissue classification. Correct tissue identification helps to improve bias-field estimation, which in turn further improves registration.

Unfortunately, several challenges have been encountered in applying iterative bias-field correction tools to heterogeneous data. First, there is a high computational cost to the iterative application of registration and bias field estimation. Second, differentiating intensity distortion from normal inter-subject variation according to tissue type is challenging. Finally, there is significant morphological inter-subject variation in the human brain, such as between subjects in different stages of disease progression, aging, or brain development.

In Wells et al. [172], they proposed an adaptive segmentation method by using the Expectation Maximization (EM) algorithm concert with bias field correction. This idea was further advanced by the work in [166, 124]. Each of these papers described an iterative method that alternates between bias correction and tissue classification within an EM algorithm framework. The paradigm has been applied to good effect on several research projects [123, 125, 124, 126]. The reported instances of these implementations, however, has only been applied with limited conditions: affine registration and/or a restricted number of tissue types.

In this paper, we expand upon the previously introduced procedure in [124], and describe algorithmic enhancements for the robustness of the framework while testing on a large-scale multi-site heterogeneous data analysis (32 sites, 3000+ scan sessions). The main improvements were achieved by incorporating deformable registration, expanding the spatial tissue definitions of 12 discrete tissue types (and 5 nuisance tissue types), and computing intensity-constrained prior based on robust *a priori* tissue-specific statistics. As a validation of this study, we compare the relative benefits of high deformable verses affine registration, 3 versus 17 prior models, and segmentation performance before versus after bias-correction.

2.2 Method

The implementation of our iterative framework incorporates 1) bias-field correction, 2) tissue classification, and 3) image registration with specific automated processing improvements for robust processing large-scale multi-site MR data. The basic philosophy of the framework has conceptual similarities to the works from Wells [172], Prastawa [124] and Tustison [7] with enhancements (dashed boxes in the Figure 2.1) that we have found useful for automated processing of large heterogeneous data.

EM Algorithm for Bias Correction: A core implementation of this work uses a general expectation-maximization (EM) algorithm [172, 166, 7, 51] for bias correction by iterating distributional parameter estimation and individual voxel y classification at location i into K tissue types. The process assumes Gaussian mixture model $y_i \sim N(\theta_i)$ where $\theta_i = \{\mu_i, \sigma_i\}$ with mean μ and variance σ^2 of each tissue label $\Gamma \in \{l | l = 1 \cdots K\}$. The first step is the expectation (E) step which determines

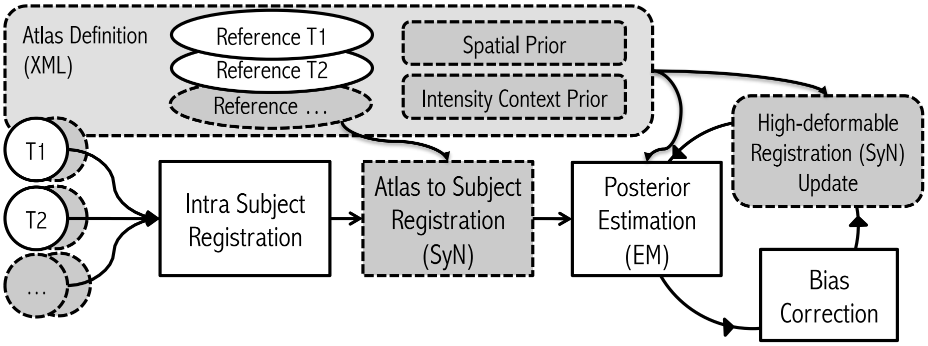

Atlas Definition (XML)
Reference T1
Reference T2
Reference ...
Spatial Prior
Intensity Context Prior
High-deformable Registration (SyN) Update
T1
T2
...
Intra Subject Registration
Atlas to Subject Registration (SyN)
Posterior Estimation (EM)
Bias Correction

Figure 2.1: **Flowchart** The framework takes any number of modalities, with any number of repetitions of scans as inputs. The algorithm starts with *Intra Subject Registration* to align all intra-session scans into first scan given. Then the initial *Atlas to Subject Registration* is performed to place all the atlas priors into subject space. Finally, the iterative process for *Posterior-Estimation*, *Bias Correction*, and *Registration Update* is repeated multiple times. Grey dashed boxes represent where our enhancements.

the expected posterior density function $p(y_i|\theta, \Phi_i)$ with estimated bias-field Φ_i [166].

(Formulations in this paper are adapted from the works [172, 166, 7, 51].)

E-Step:

$$p(y_i|\theta, \Phi_i) = \sum_l p(y_i|\Gamma_i = l, \theta_l, \Phi_i)p(\Gamma_i = l), \tag{2.1}$$

with $p(y_i|\Gamma_i = l, \theta_l, \Phi_i) = N_{\sigma,l}(y_i - \mu_l - \Phi_i)$ and

$$p(\Gamma_i = l) = \frac{t_{il}}{\sum_{l=1}^{k} t_{il}}, \text{ where } t_{il} \text{ is a tissue specific prior} \tag{2.2}$$

Next, the maximization (M) step computes parameters of Gaussian θ and bias-field Φ by maximum likelihood estimation from the current density function.

M-Step:

$$\mu_l = \frac{\sum_i p(\Gamma_i = l|y_i, \theta, \Phi_i)(y_i - \Phi_i)}{\sum_i p(\Gamma_i = l|y_i, \theta, \Phi_i)} \tag{2.3}$$

$$\sigma_l^2 = \frac{\sum_i p(\Gamma_i = l|y_i, \theta, \Phi_i)(y_i - \mu_l - \Phi_i)^2}{\sum_i p(\Gamma_i = l|y_i, \theta, \Phi_i)} \tag{2.4}$$

The previous equations are extended to multi-modal data as described in [166].

2.2.1 Overview of Proposed Procedure with Enhancements:

This tool begins by taking a collective set of multi-modal MR images as input with any number of repetitions. Repeated scans within a single sessions can be taken advantage of to increase the signal-to-noise (SNR) ratio for each modality. Our procedure begins with spatial normalization of each intra-modal scan into a common subject-specific reference orientation defined by the AC (anterior commissure), PC (posterior commissure), and mid-sagittal plane by using a Rigid-type transformation [54]. The spatial normalization reduces non-subject specific spatial variation between scans, and in turn, enhances robustness and efficiency of subsequent steps. Next, subject-specific tissue posteriors are estimated by performing the EM procedure described previously. The posterior estimation step here employs 1) an atlas-to-subject high-deformable diffeomorphic registration algorithm (ANTS [7]) enhancing accuracy of subject specific tissue priors by increasing warping correspondence to the subject and 2) a novel region-specific intensity constraint to ensure the correctness of tissue posteriors. After the E-Step, the bias-field of each input MR image is estimated in the M-Step and applied based on current estimate of tissues. The iterative process goes back to posterior estimation using the improved intensity homogeneity, consecutively improving the estimation of the inter-scan registration, and so improves the tissue posterior estimation.

2.2.2 A collective set of input: Multi-modal MR Images with repetition

A collective set of multi-modal MR images including repetitions from a single scan session are utilized. It is well established that multi-modal MR images can provide collaborative knowledge that can improve brain tissue classification [139]. We further employed repeated scans, when they were available, to increase the SNR for those modalities. Through careful study design and scanning protocols can limit the occurrence of artifacts, some are unavoidable. By taking repeated scans in a single scan session, artifacts such as noise can be reduced sufficiently.

2.2.3 Integrating high-deformable Registration [SyN [8]]

High deformable registration is integrated for accurate deformation mapping estimation between atlas priors and subject-specific space. We hypothesize that improved correspondence of the atlas to the subject benefits tissue classification as well as bias-field correction compared to previously employed affine or B-Spline registrations. Symmetric image normalization (SyN) based registration [8] provided from the ANTS package is extensively tested and has been shown to perform well at preserving image topology. With this highly deformable registration ψ, our subject specific spatial priors are now further refined:

$$p(\Gamma_i = l) = \frac{\psi(t_{il})}{\sum_{j=1}^{K} \psi(t_{ij})} \tag{2.5}$$

2.2.4 Extended Prior Definition

Traditionally, spatial priors are used to intelligently propagate tissue-specific spatial knowledge to a MR image-processing algorithm. One of the main assumptions behind the utilization of tissue spatial priors is the homogeneous intensity profile of identical tissue type across images of the same modality. In a large-scale study setting, however, this is violated and the degree of inhomogeneity can be vastly different from scan to scan. Rather than adjusting the algorithm's parameters for each problematic case, the pragmatic way to deal with the situation is to break down the biological tissue definitions further by their unique image properties. We identified 17 extended tissue specific priors based on their spatial location, intensity profiles, and biological definitions (Figure 2.2). The priors are constructed to have intrinsic hierarchical tissue definitions with respect to each other (See Table 2.1 and Figure 2.2)

The most prominent regions of interest include grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). These regions of interest are further partitioned based on their spatial property: depending on whether it is located in the cortical/subcortical (peripheral/central) area of brain or outside of the brain. The distinction between cortical and subcortical tissue is practically useful since they often present heterogeneous MR intensity characteristics across different imaging modalities. Tissues outside the brain are named 'not-tissue' regions, such that they demonstrate similar MR-image profiles to tissues of interest but are spatially outside of the brain, such as bone, fat, or skin. Designation of each spatial tissue prior is summarized in Table 2.1.

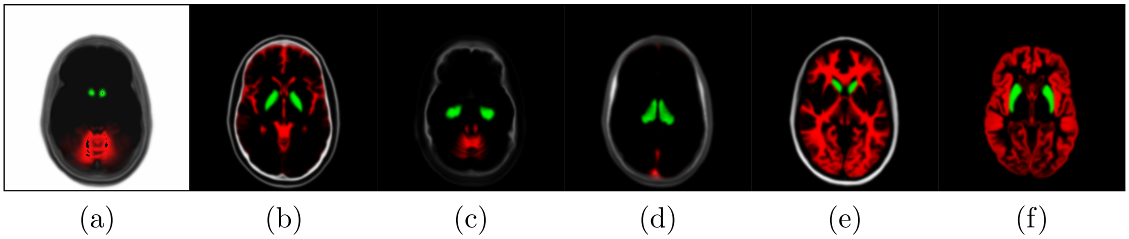


Figure 2.2: 17 Spatial Priors. *(Denote grey [gr], green [gn], and red [r])* (a): air [gr], nucleus accumben [gn], and cerebellum GM [r], (b): not-GM [gr], globus pallidus [gn], and CSF [r], (c): not-CSF [gr], hippocampus [gn], and cerebellum WM [r], (d): not-venous blood [gr], thalamus [gn], and venous blood [r], (e): not-WM [gr], caudate [gn], and cerebral WM [r], and (f): putamen [gn] and surface cerebral GM [r].

Table 2.1: Atlas Definition of 17 Region-Specific Intensity-Context Prior

Tissue	Name	Weight	Bias correction	q^{T1}_{lower}	q^{T1}_{upper}	q^{T2}_{lower}	q^{T2}_{upper}
GM	Accumben	1	-	0.05	0.95	0.15	0.97
	Caudate	1	-	0.05	0.95	0.15	0.97
	Crbl Gm	1	True	0.03	0.9	0.02	0.99
	Hippocampus	1	-	0.05	0.95	0.15	0.97
	Putamen	1	-	0.05	0.95	0.15	0.97
	Surf Gm	1	True	0.04	0.75	0.25	0.96
Wm & Gm	Thalamus	1	-	0.05	0.95	0.15	0.97
	Globus	1	-	0.05	0.95	0.15	0.97
WM	Wm	1	True	0.5	1	0.05	0.7
	Crbl Wm	1.5	True	0.1	1	0.03	0.9
Csf	Csf	1	True	0	0.6	0.2	1
VB	Vb	1	-	0.04	0.75	0	0.2
Bg	Not Csf	1	-	0	0.6	0.2	1
	Not Gm	1	-	0.15	0.9	0.35	1
	Not Vb	1	-	0.15	0.9	0	0.3
	Not Wm	1	-	0.4	1	0.1	0.85
	Air	1	-	0	0.1	0	0.1

Each *Tissue* type sub-divided into regions of interest with given *Name*. *Weight* and whether it is used for *bias correction* computation are also shown. Intensity-context prior definitions are their valid range according to quantized percentile of intensity, Quantiles ($q^{T1}_{lower/upper}, q^{T2}_{lower/upper}$), to take account MR's relative intensity and outliers.

2.2.5 Multi-modal Region-Specific Intensity-Context Priors

Intensity-context priors are devised for algorithmic robustness of large-data processing. With our enhanced prior-definitions, there were still some cases that failed due to false positives of tissues. These threshold-identified regions were used as an additional constraint in conjunction with their corresponding spatial priors. Since MR intensities are are not a standardized quantitative measurement, the threshold parameters for each tissue type were designated by quantiles of each images histogram. The multi-modal quantile threshold value of each tissue type was conservatively chosen to ensure that each tissue regions was completely included (i.e. no false negatives). The set of multi-model threshold parameters used globally for our studies are shown in Table. 2.1. By incorporating *a priori* knowledge β, i.e., the multi-modal intensity constraints of each tissue type, the initial estimates of tissue statistics are made more robust across a wide range of imaging protocols. Therefore, for multi-modal intensity model $\bar{y}$ with a indicator function, $\bar{\mathbf{1}}_\beta(\bar{y}_{il})$, Eq. 2.2 and Eq. 2.5 can be further refined:

$$p(\Gamma_i = l) = \frac{\psi(t_{il}) \cdot \bar{\mathbf{1}}_\beta(\bar{y}_{il})}{\sum_{j=1}^{K} \psi(t_{ij}) \cdot \bar{\mathbf{1}}_\beta(\bar{y}_{ij})}, \tag{2.6}$$

$$\bar{\mathbf{1}}_\beta(\bar{y}_{il}) = \begin{cases} 1 & , \text{ if } \bar{y}_i \in \beta = \{\bar{y} | q^m_{lower} < y^m < q^m_{upper} \text{ for all image } m \ \} \\ 0 & , \ otherwise \end{cases}$$

2.3 Evaluation

The accuracy and effectiveness of our proposed enhancements were evaluated from multiple perspectives: 1) compared similarity against known ground truth by using simulated MR data, 2) visual inspections of results by experts, and 3) a visual comparison of sample results to well-established techniques (FreeSurfer [47, 134] and Atropos from ANTS package [7]).

2.3.1 Evaluation using Simulated Data

A series of evaluation experiments using BrainWeb data with six levels of simulated noise, two degrees of bias-field are summarized in the Figure 2.3.

Two independent measures, Dice index (larger is better) and average Hausdorff distance (smaller is better), are reported to underscore the validity of our processing between automated delineation and ground truth. The visualization results make it clear there is a difference between segmentation with and without enhancements. Along the six noise levels with two degrees of bias-field, the most agreeable result to the ground truth was obtained by utilizing all of our proposed enhancements (Figure 2.3:**black**).

The series of BRAINWeb experiments demonstrate the benefit of each individual enhancement, and also the combined benefit of using all enhancements together. High deformable SyN registration improves tissue classification results as compared to affine registration (**blue** versus **black**). Second, multi-modal intensity constraints improve the procedure, especially when registration is less than optimal due to large morphological differences, often present in degenerative diseases such as HD (Figure 2.3: contrasting **yellow** versus **blue**). Third, the extended definition of tissue priors helps to increase the segmentation accuracy (Figure 2.3:**purple** versus **black**). These three improvements are valid for all six levels of noise and both bias-field levels.

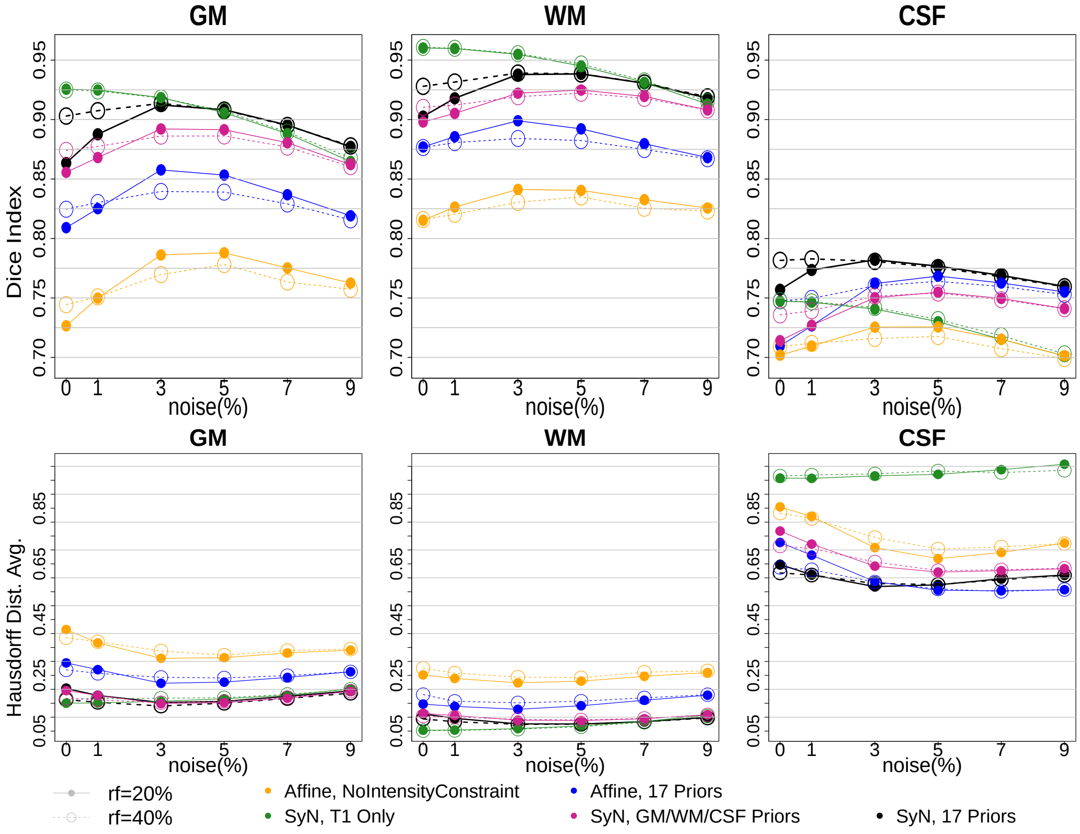


Figure 2.3: Two bias-fields, *rf=20* (solid line) and *rf=40* (dotted line), are shown with six noise levels along x-axis. Two independent measures of Dice Index (upper) and average Hausdorff Distance (bottom) are shown. With affine registration, *Intensity-context prior* (yellow) has better accuracy than one without one (blue). *SyN* registration (pink, black) also improved tissue segmentation agreement further comparing to the *affine* (yellow, blue). Note that the performance with *SyN* registration utilizing only three tissue types outperformed any of affine-based method. *Multi-modal input trial*, T1 and T2, (black) comparing to *T1 Only* (green) seems helpful when there is more noise.

Finally, using multi-modal input is beneficial, especially when MR scan is corrupted with noise and/or inhomogeneity bias (Figure 2.3: **black** versus **green**).

2.3.2 Evaluation on In-Vivo Data

The proposed pipeline was applied on in-vivo MR data, collected from the multi-site international PREDICT-HD [127] project. The PREDICT-HD data [127] is highly heterogeneous. The inhomogeneity of the data is due to the multi-site natural history observational study design that employed all the available resources, including multiple MR vendors (GE, Phillips, and Siemens), field strengths ($1.5T$ and $3T$), and over 20 different MR acquisition protocols (i.e., due to transmission and receive hardware). All the processed images (n=3751) are visually inspected by three independent experts and only $< 2\%$ scans were classified as failing to produce bias-corrected T1 images. Processing time varied approximately between 3-5 hours per scan for the entire bias correction, registration, and tissue classification and was related to the number of modalities and repeats in the scan.

Three subjects were sampled to retrospectively compare results with regard to MR vendors and the rough estimate of WM/GM tissue ratio (volume of WM and GM over intracranial volume) to reflect expected HD-specific morphological variations in our large-scale multi-site study. Characteristic of the sample data is summarized in Table 2.2.

The smaller tissue ratio means more atrophy in brain tissue, which is generally caused either by disease progression or aging. To facilitate visual comparison to

Table 2.2: Sample results are shown in this paper for visual comparison

scan	site	MR Vendor	Field Strength	Tissue Ratio	Collected Modality(#)
A	Site_180	SIEMENS TrioTim	3T	> 0.88	T1(2), T2(2)
B	Site_024	GE	1.5T	< 0.78	T1(1)
C	Site_039	PHILIPS	3T	< 0.78	T1(4), T2(2)

Sample results are shown in this paper for visual comparison: processing results from three subjects that are collected at different sites with various characteristics in MR vendors, rough tissue ratios, field strengths, and number of repeats.

other works, tissue classification results from Atropos [7] and FreeSurfer [47, 134] are displayed in Figure 2.4, 2.5, and 2.6 To be equivalent, multi-modal approaches (where applicable) were applied and their results visually inspected for all three processing pipelines. As the figures show, tissue boundaries in the cortical area (peripheral region of brain) from our proposed approach (*BRAINSABC*) are more agreeable to other two methods than subcortical area (central area of brain). For the subcortical GM, however, the three approaches resulted in noticeable differences. *BRAINSABC*'s results were closer to FreeSurfer while Atropos exhibited the most conservative subcortical GM delineation. Note also that the globus pallidus was treated as an independent tissue type in BRAINSABC and FreeSurfer while there was no special consideration for globus pallidus in Atropos.

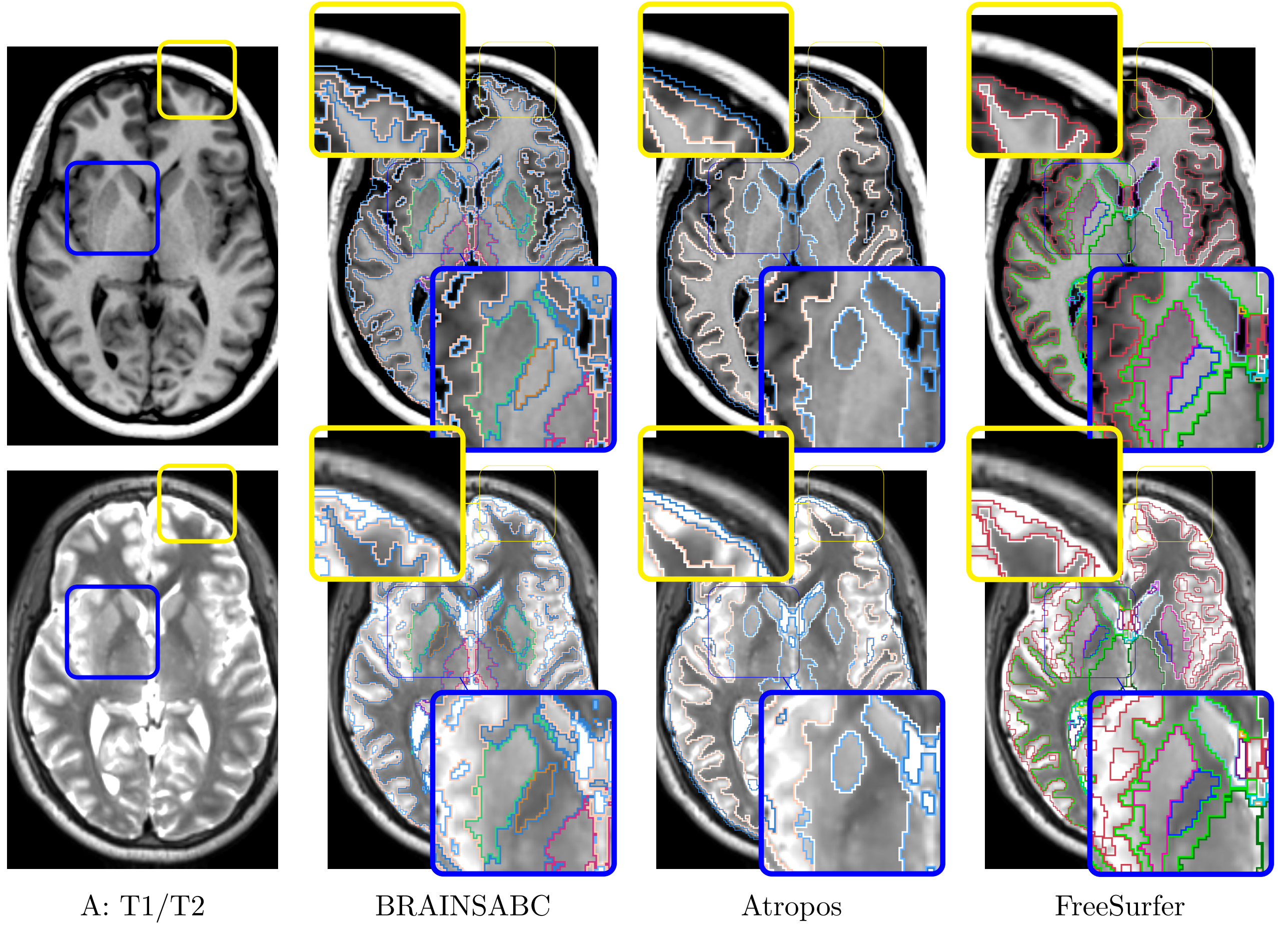


Figure 2.4: Visual Comparison of Scan A in Table 2.2. Tissue classification results from BRAINSABC, Atropos of ANTS Tools, and FreeSurfer on top of T1- and T2-weighted. This scan has relatively normal tissue ratio, reflecting minimal, if any, atrophy in brain tissues. Yellow and blue boxes highlight where tissue classification is more differentiated from each other. BRAINSABC was highly agreeable to Atropos and FreeSurfer for the cortical area in general. For the subcortical area (blue box), however, BRAINSABC produced more agreeable to FreeSurfer than Atropos, which was the most conservative in GM identification. For the CSF, which is more distinguishable in the T2-weighted modality, BRAINSABC produced robust results.

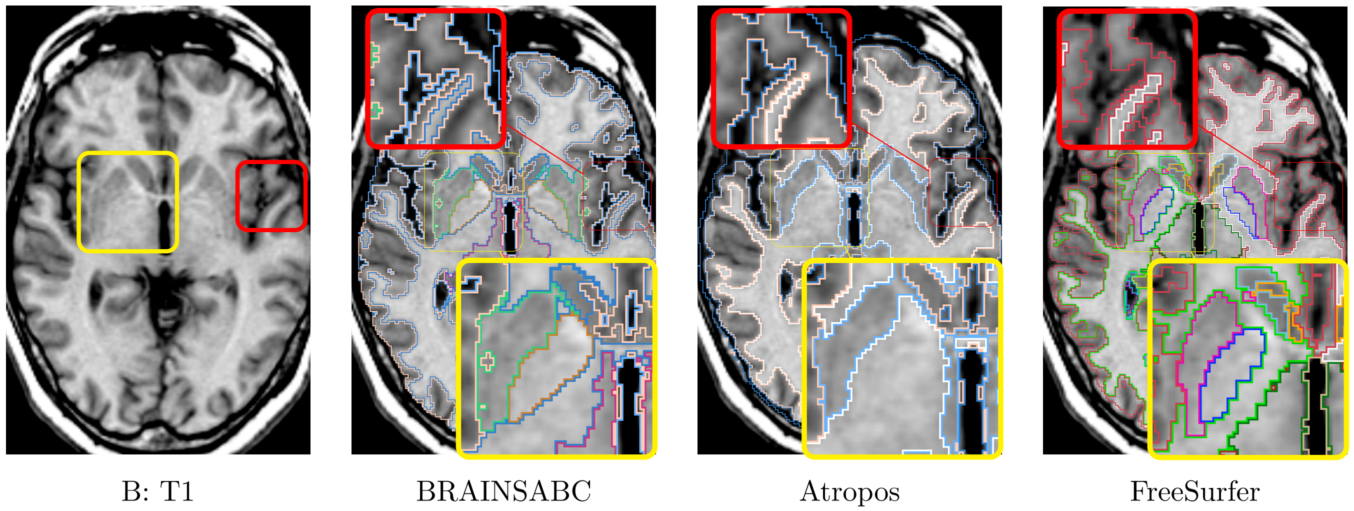


Figure 2.5: Visual Comparison of Scan B in Table 2.2. Tissue classification results from three applications, BRAINSABC, Atropos of ANTS Tools, and FreeSurfer on T1-weighted images using uni-modal processing. This subject presented a relatively small tissue ratio, reflecting advanced brain atrophy progression. Again, Atropos was more cautious in the subcortical (yellow box) GM identification than others. Red box also underlines differences of tissue classifications on the cortical area. Note that without T2 modality, CSF classification results of BRAINSABC were more agreeable to the other two.

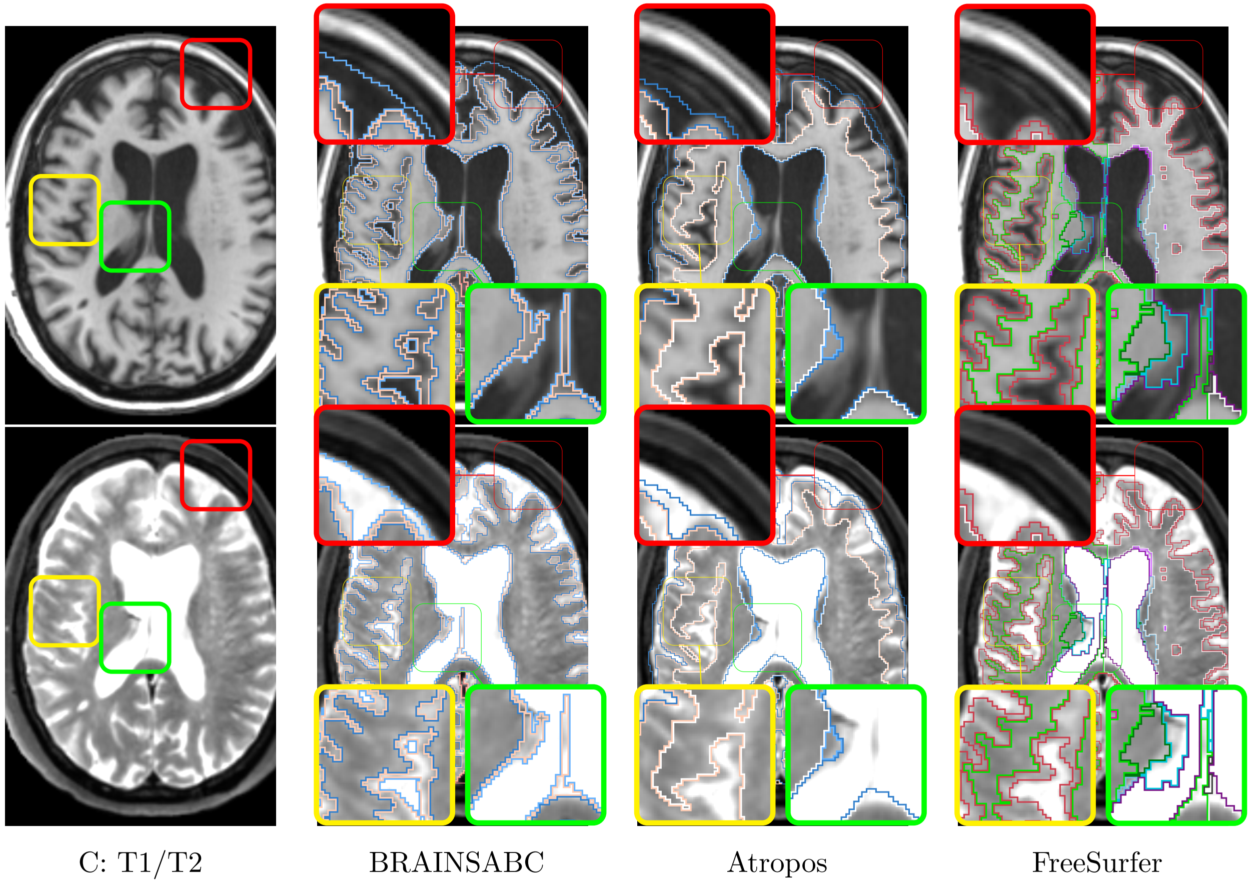


Figure 2.6: Visual Comparison of Scan C in Table 2.2. Tissue classification results from BRAINSABC, Atropos of ANTS Tools, and FreeSurfer on T1- and T2-weighted images. This subject presented a relatively small tissue ratio, reflecting brain atrophy progression. All produced very similar results in this scan. Red box contrasts classification on the cortical, where BRAINSCut draws a nice borderline for surface CSF. Green box magnified CSF and GM border on the subject with enlarged ventricles where T1-weighted image shows artifacts. Yellow box also shows different classification results between tools. Again CSF classification presents some disagreement between FreeSurfer and BRAINSABC (green box).

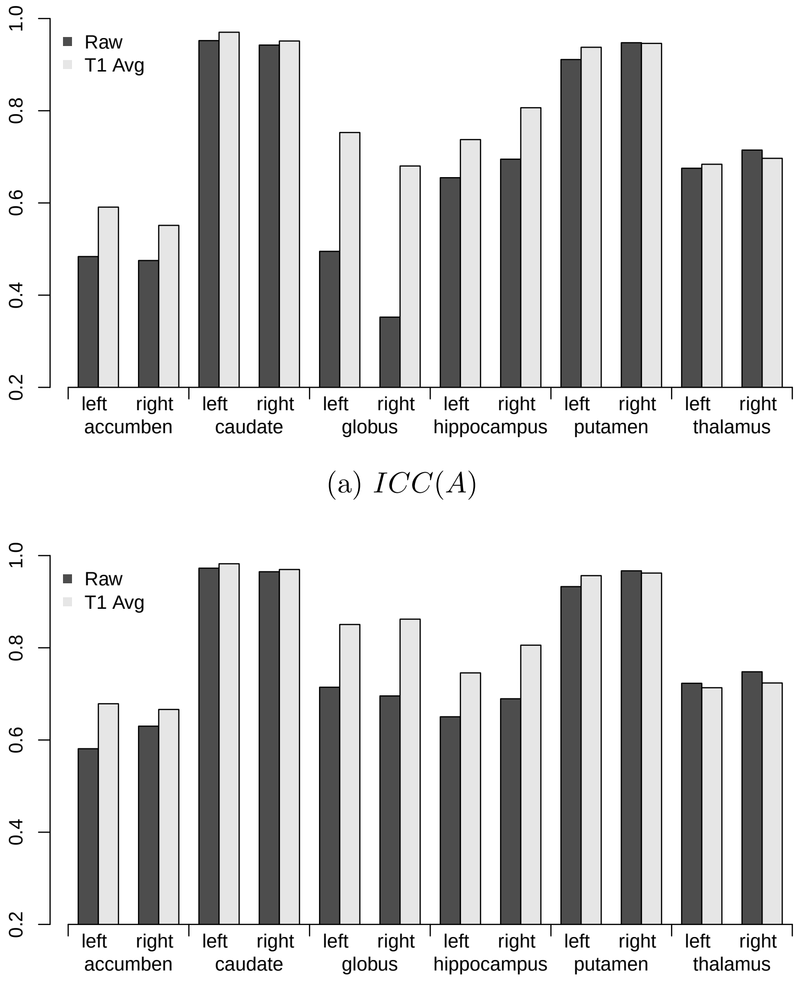
2.3.3 Evaluation using Contribution to the Segmentation Accuracy

The benefits of the bias-corrected MR images toward segmentation accuracy is apparent in Figure2.7 shows. Segmentation accuracy of all subcortical structures but thalamus were effectively improved bias-corrected inputs. Thalamus shows negligible decrease in ICC measures when bias-field corrected inputs are employed.

2.4 Discussion

BRAINSABC, the proposed bias-field corrector, has effectively been improved for large-scale in-vivo data analysis. The key contributions of this work are four fold: 1) pipeline enhancements for large-scale heterogeneous MR data processing, 2) empirically showing advantages of utilizing multiple scans including multi-modal or repeated MRIs, 3) distributing all the tools of the open source pipeline including all parameter sets, and 4) explicitly proving the advantage of bias corrected input for the subcortical segmentation.

Additional advantages of our proposed enhancements are revealed from in-vivo application. First, BRAINSABC requires no pre-alignment between scans because the process incorporates both intra-subject and atlas-to-subject registration with refinements in the iterative process. Minimal pre-requisite of softwares lead to a more applicable for any clinical studies. Second, as shown in Figure 2.6, 2.5, and **??** the brain extraction of BRAINSABC produces very high quality brain region estimate as compared to other two approaches. A robust brain region estimation is important because it is often employed in normalizing sub-volumetric data to compensate



(a) $ICC(A)$

(b) $ICC(C)$

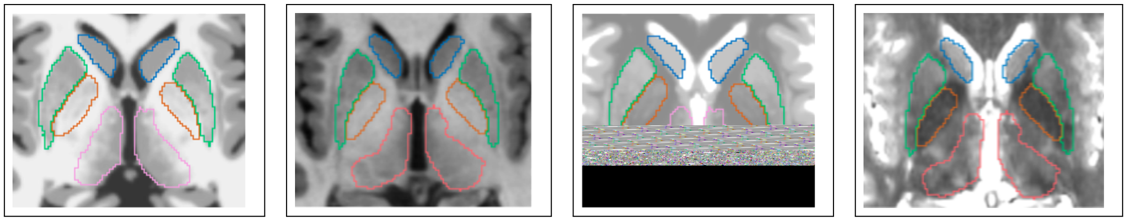
Figure 2.7: Segmentation accuracy is contrasted before ($T1Avg$) and after (Raw) bias-corrected T1 images. ICCs of six subcortical segmentations between automated method and manual traces from 10-fold cross-validation experiments. Improvement of segementation accuracy is apparent in terms of both ICCs of agreement ($ICC(A)$) and consistency ($ICC(C)$). Note that interquantile range (IQR) based region-specific intensity normalization is used for all structures but the caudate nucleus; $linear(min/max)$ transform based on min/max value were used for caudate nucleus. The details about the choice of normalization approaches appear the independent research (Chapter 4).

for overall brain size differences between subjects. In addition, we found empirically that the visual inspection failure rate of FreeSurfer on the raw MRI scans was approximately 20%, on large-scale heterogeneous MR data, but when FreeSurfer was provided the BRAINSABC tissue classified pre-aligned and bias-corrected images the visual inspection failure rate dropped to approximately 8%.

With the evaluation on the simulated data in Section 2.3.1, Dice similarity coefficient (DSC) goes up at first and then down as the noise level increases. One possible explanation for the slight DSC increment as a bit of noise/bias added to a simulated MRI, is a difference between MRI without noise and in-vivo MRI (See Figure 2.8). A patient MRI in-vivo is usually corrupted by noise to some extent. Since our techniques are highly optimized for in-vivo MRIs, testing the method on simulated MR images without realistic noise corruption, may be a less than ideal validation.

A more comprehensive evaluation and validation of all available tissue classification tools to see how our proposed tool performs in comparison would have been ideal, but this task was determined beyond the scope of this paper. However, in this study, we have provided a formal validation study of our proposed tool as well as a formal comparative study against well-established tools. In addition, the results of our study has undergone a rigorous qualitative assessment that involved visual inspections by three independent experts who have been trained on a large number of scan sessions from various sites and scanning protocols.

The software implementation is written based on the *InsightToolkit* libraries and conforms to the coding style, testing, and software license guidelines specified by the National Alliance for Medical Image Computing group. Our implementation, BRAINSABC, is publicly available at `https://github.com/BRAINSia/BRAINSTools` via BRAINSTool package. Our implementation is optimized for, but not limited to, large-scale MR data analysis. The implementation has successfully applied over 3000 scans from the large-scale longitudinal study [127] and visually inspected their validity. Advantages of our tool come primarily from its demonstrated generalizability to a wide number of scanning protocols, variations in the number and type of modalities, and number of repeated scans. As a part of larger image processing framework, this iterative automatic bias-field correction module provides very robust and consistent results for further MR image analysis.



(a) BrainWeb-T1 (b) Real-T1 (c) BrainWeb-T2 (d) Real-T2

Figure 2.8: Distinguishable intensity profile differences between BrainWeb and real MR data. Difference is more obvious for the globus pallidus (orange) region, where BrainWeb data appears medium gray both in T1 and T2. Real data, however, presents the globus pallidus (orange) with medium gray in T1 and dark gray in T2.

2.4.1 Summary & Conclusion

We described a method to improve the performance of automatic bias-field correction algorithm for large-scale heterogeneous MR data processing. The excellent robustness of the tool against large-scale clinical data was achieved by collaborating multiple interdependent enhancements in the iterative process. Our proposed enhancements are evaluated via application to both the simulated brain MR images as well as in-vivo MRI collected from a large multi-center study. The series of experiments on simulated MR data revealed the improved robustness by our proposed enhancements in the presence of varying levels of noise, and inhomogeneity. In addition, application to in-vivo MRI, collected for multicenter study, also showed good generalizability as demonstrated by the very low failure rate across a wide spectrum of input image protocols. Sample results are also confirmed a competitive performance of our tool in comparison to the well-established tools of Atropos and FreeSurfer. The bias-corrected images from the *BRAINSABC* also generally improved the subcortical segmentation accuracy.

CHAPTER 3
OPTIMAL MACHINE-LEARNING ALGORITHM SELECTION

This chapter describes a robust machine-learning (ML) algorithm selection for segmentation framework in order to process a large amount of scalable MRI data collected from different centers and institutions. Our segmentation framework was constructed to utilize a machine-learning algorithm. This chapter conducts a comprehensive investigation to find the best one from several available algorithms in the field. Experiments are designed hierarchically to effectively filter out inferior ML algorithms. First, a screening study compared twelve various machine-learning algorithms and identified the artificial neural network (ANN) and the random forest algorithm as the most promising. The second phase of experiments focused on contrasting two most strong candidates, ANN and random forest algorithm, and revealed the superiority of random forest algorithm for the subcortical MRI segmentation framework.

3.1 Introduction

Precise delineation of sub-cortical structures from structural magnetic resonance images (MRI) is advance the understanding of the human brain. A rich set of MRI soft tissue information [139] provides opportunities for quantitative studies detecting subtle morphological changes during disease progression. With advanced MRI techniques, trajectories of neuroanatomical morphology in normal aging is are now known to differ from neurological diseases such as schizophrenia, Alzheimer's disease, autism, Huntington's disease, and others [153, 66, 163, 10, 43, 11, 113, 40, 9, 115, 52, 107, 149, 114, 45, 12, 108, 13]. Recently, large-scale MRI data collection with multicenter collaborations ([127, 161]) has been conducted to obtain more sensitive models of pathological changes in brain structures. Efficient and robust delineation methods for scalable data is therefore critical to deepen our understanding of neurological disease trajectories.

Difficulties, however, exist in conducting robust quantitative assessments on MRI data. While manual tracing remains common practice for anatomical delineation, this solutions are limited by the high cost of manual labor and low intra- and inter-rater consistency. This is especially true when it comes to large-scale longitudinal multicenter studies, where the manual approach prevents efficient quantitative MRI analysis.

Several automated tools have been developed [14] to address the drawbacks of manual traces. Automated segmentation tool development, however, has multiple challenges as well. First, the tool must be able to handle large data variation given a limited amount of prior information. Human experts delineate MRIs with dynamic inference based on empirical knowledge and pre-attained information from biology, physiology, neurology, and anatomy. In contrast, the automated computer algorithm is constrained by restricted static information – usually voxel intensity and location. Secondly, intra-scan intensity inhomogeneity, one of the inevitable obstacles in MRI, poses a primary challenge for a successful automated tool by corrupting that little information is available to it. Third, for large-scale multicenter MRI data the intensity profile variability is highly heterogeneous compared to one from a traditional single-site study. These variations arise from the diversities in manufacturers, protocols, and the field strengths of the MR scanners. With these challenges, extracting reliable volumetric information from MRIs is still an active area of research.

There have been very few attempts to develop an automated tool for processing large-scale multicenter MR data. Among various available methodologies, machine-learning based segmentation algorithms are attractive because they are generalizable and robust for scalable data analysis. A number of machine-learning techniques have been employed for automated segmentation in the literature: SVM [181, 2, 60, 138, 104], AdaBoost [104], k-NN [168, 4], and ANN [122, 84]. There is, however, little research that provides a broad comparative study on the relative benefits of each machine learning approach within a brain segmentation framework.

Two contributions are presented in this chapter. First, we conducted a screening study which contrasted twelve variations of the eight machine-learning algorithms on identical MR segmentation data to choose the optimal candidate algorithms. Second, based on the screening study we exhaustively explored the two most promising methods, (random forest and ANN), within the established segmentation framework [122, 84].

3.2 Background

This study focuses on developing a reliable and efficient machine-learning (ML)-based segmentation tool that provides results that highly corresponds to human tracers. We describe shared properties among machine-learning algorithms (Section 3.2.1), algorithm-specific details used in this study (Section 3.2.2 and Sec 3.2.3) and finally related research (Section 3.2.4).

3.2.1 Machine-learning Model: Common Properties

> *Learning denotes changes in a system that enable a system to do the same task more efficiently the next time.* *(Herbert A. Simon, 1983)*

In our framework, the machine-learning classifier is the central processing step for the region segmentation. We first describes shared properties between machine-learning algorithms (Section 3.2.1.1) and common issues raised in machine-learning method (Section 3.2.1.2).

3.2.1.1 Goal of Machine-learning Model

A successful machine-learning model $\mathbb{M}$ is to be trained to better fit the example data $\mathcal{S}$, and thus $\mathbb{M}$ becomes more applicable to a larger set of real-world data $\mathcal{X}$. For a given pattern f_i, a machine-learning algorithm *predicts/estimates* a label $\hat{y}_i$ of each voxel at a location i. The model $\mathbb{M}$ is built through a *training phase* on the given example set $\mathcal{S}$.

The fit of the model $\mathbb{M}$ can be estimated by the total misclassification rate for $\mathcal{S}$. The total misclassification rate, also commonly called *error*, is generally used to assess the performance of the trained model $\mathbb{M}$. There exists two different kinds of error involved in machine-learning development:

1. *True prediction error* [e]: the true underlying error distribution on the research population data $\mathcal{X}$. Since true error e is unmeasurable, it is usually replaced by an estimated value e'.

2. *Apparent error* $[e_a]$ (Training error, re-substitution error): error rate that is calculated from the training data $\mathcal{S}$. Apparent error could be used as the estimator e' for true error, at the expense of underestimation of e. e_a is usually biased, and called *wildly optimistic* of the true error ($e' = e_a \ll e$). Apparent error is defined as

$$e_a = \sum_{\forall i \in \mathbb{I}} Err(y_i, \hat{\mathbb{M}}(f_i)),$$

where $Err()$ is a function to compute differences between known label Y and the estimated label $\hat{\mathbb{M}}(\mathbf{F})$.

Another common mean assessing performance of machine-learning model the is *confusion matrix*. The confusion matrix is a specific table visualizing performance of a machine-learning algorithm [128]. The confusion matrix provides information about the desired output and the estimated label from ML classifier (Tab. 3.1) and various performance metrics can be computed from this matrix:

Table 3.1: Confusion Matrix

		Predicted	
		negative	positive
actual	negative	TN	FP
	positive	FN	TP

Confusion Matrix. A table visualizing performance of a machine-learning algorithm. True negative (TN), false positive (FP), false negative (FN), and true positive (TP).

$$\text{Accuracy : } AC = \frac{TN + TP}{Total} \tag{3.1}$$

$$\text{True positive rate : } TPR = \frac{TP}{(TP + FN)} \tag{3.2}$$

$$\text{False positive rate: } FPR = \frac{FP}{(FP + TN)} \tag{3.3}$$

$$\text{True negative rate : } TNR = \frac{TN}{(TN + FP)} \tag{3.4}$$

$$\text{False negative rate: } FNR = \frac{FN}{(TP + FN)} \tag{3.5}$$

$$\text{Precision : } P = \frac{TP}{(TP + FP)} \tag{3.6}$$

$$\text{Sensitivity, Recall: } Rc = \frac{TP}{(TP + FN)} \tag{3.7}$$

$$\text{Specificity : } Spec = \frac{TN}{(FP + TN)} \tag{3.8}$$

Note that when assessing and developing a machine-learning prediction model, being able to accurately measure its *prediction error* is of foremost importance. Correct estimation of the prediction error can lead to the building an accurate prediction model, while on the other hand, the use of an incorrect error measure can result in the selection of an inferior and inaccurate model $\mathbb{M}$.

3.2.1.1.1 Bias and Variance

Apparent error e_a is often decomposed to two components, *bias* and *variance*, to better understand behavior of trained machine-learning models. Bias measures how far off from the mean of Y the those predictions of $(\hat{\mathbb{M}}(f))$ are and variance measures how much the predictions for a given point vary between different realization of the model. Mathematically, the true misclassification rate is an expected difference between the true (unknown) label Y and estimated label by $\hat{\mathbb{M}}$ given feature f.

$$e(f) = E[(Y - \hat{\mathbb{M}}(f))^2].$$

Then, the prediction error can be partitioned into two subcomponents: bias and variance.

$$e(f) = \left(E[\hat{\mathbb{M}}(f)] - \mathbb{M}(f)\right)^2 + E\left[\hat{\mathbb{M}}(f) - E[\hat{\mathbb{M}}(f)]\right]^2 + \sigma_N^2$$

$$e(f) = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

$$Bias : E[\hat{\mathbb{M}}(f)] - \mathbb{M}(f)$$

$$Variance : E\left[\hat{\mathbb{M}}(f) - E[\hat{\mathbb{M}}(f)]\right]^2$$

The above equations present a mathematical *bias-variance decomposition* for formal assessments of the prediction error e of a machine-learning model $\hat{\mathbb{M}}$.

Given the true model $\mathbb{M}$ and infinite sample ($n \to \infty$) to calibrate it, we should be able to reduce both the bias and variance terms to zero. However, in a world with imperfect models $\hat{\mathbb{M}}$ and finite data ($\mathbf{S}_n, n < \infty$), it is common to trade-off some increase in the bias for a larger decrease in the variance and vice-versa [89].

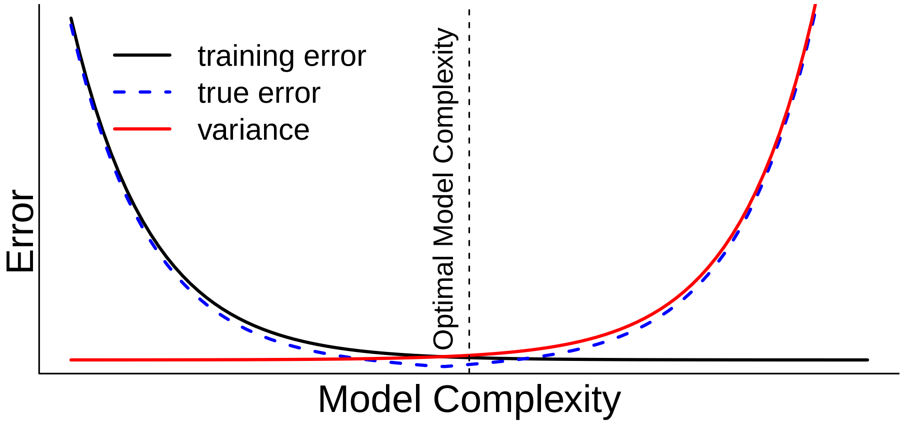


Figure 3.1: **Error Terms in relation to model complexity for machine learning algorithm.** True error e (black) is presented with *variance* (red) and *bias*^{2} (black) along the model complexity. As model complexity increases, the variance also increases while *bias*^{2} drops rapidly. Due to the *variance*, the total error has a natural minimum and this is the point of the optimal model complexity. After the optimal model complexity, a machine-learning model is regarded as having *overfitted* the training data. Overfitting issue is also described in Section 3.2.1.2.

3.2.1.2 Overfitting

In the previous section (3.2.1.1), we showed that the variance exponentially increases as the model complexity increases in Figure 3.1. This is called *overfitting* of the estimated model for the specific training data **S**. That is, the trained model poorly performs outside sample $\mathcal{S}^{\mathsf{C}} \subset \mathcal{X}$, while presenting great accuracy on the given example $\mathcal{S} \subset \mathcal{X}$, where $\mathcal{S}^{\mathsf{C}} \cup \mathcal{S} = \mathcal{X}$. This happens when the model fits only the pattern of the given sample $\mathcal{S}$ through minimization of the apparent errore_a. Ideally, a machine-learning model should be applicable to the larger, unseen real-world data by avoiding overfitting.

Cross-validation is one way to avoid overfitting by accurately estimating the population error e. It is known that the apparent error as an estimator e' for true error ($e' = e_a$), generally under-estimates the population error, $e' = e_a << e$. On the other hand, estimated true error from cross-validation $e' = e_c$, provides a nearly unbiased estimator [110]. With k-fold cross-validation, the entire sample is randomly divided into k subsets, and of k sets, a single subset retained for testing while $k - 1$ subsets are used for training. The cross-validation then repeats this process until each subset participated as a testing set. Then k errors are averaged over to calculate the cross-validation error e_c

$$e' = e_c = (1/k) \sum e^k$$

The advantage of k-fold cross-validation is that the error term is computed from the unseen data, thus providing a significantly unbiased estimator for e.

In addition to cross-validation, overfitting can be avoided by reducing **model complexity**. Occam's Razor provides sage advice about model complexity:

> ***Occam's Razor:*** *Entities should not be multiplied beyond necessity.*
> *Every theory should be exactly as complex as necessary but no more so.*

In machine learning, overly complex models can result in overfitting of the training data such that the resulting model is no longer valid in a more general application. Since the strategies of reducing model complexity depends on the ML algorithm in question, this is discussed for each algorithm in the following section.

3.2.2 Generic Machine learning algorithms

Eight general machine learning algorithms are described here. We describe how each method works, what theoretical and practical properties are associated with, and what possible limitations exist. There are several references for each method so we take our description mostly from [110].

3.2.2.1 Majority Classifier

A majority classifier simply predicts all the instances by the majority found in the training data. If there is no obvious majority, it can be determined arbitrarily. Although not predictable, majority classifiers often serve as a moderate lower bound (baseline) to a comparative study.

3.2.2.2 Naïve Bayes

Naïve Bayes infers the outcome by estimating the posterior based on Bayes theorem:

$$P(L|F) = \frac{P(L \cap F)}{P(F)} = \frac{P(L)\prod_{f \in F} P(f|L)}{P(F)}. \tag{3.9}$$

where L is a set of classes or outputs, and F is a corresponding feature or input vector. Bayes theorem replaces the often *difficult* calculation of 'posterior' to the *easier* computation of two 'priors' and one conditional probability. The strength of Naïve Bayes in practice is to isolate noise and irrelevant features in F because such data is averaged out when estimating the conditional probabilities [110]. As we see from the equation (3.9), however, the joint probability calculation ($P(L \cap F)$) assumes statistical independence between all predictors in F:

$$P(L \cap F) = P(L) \prod_{f \in F} P(f|L) \tag{3.10}$$

As a result, predictability (the classifier's accuracy) is susceptible to dependencies between input features (e.g., f_a and f_b, $f_i \in F$). This independent assumption is often not valid in practice. Indeed, in our case it is violated because of the redundancy of scan modalities we chose in order to suppress the image noise in the MRIs.

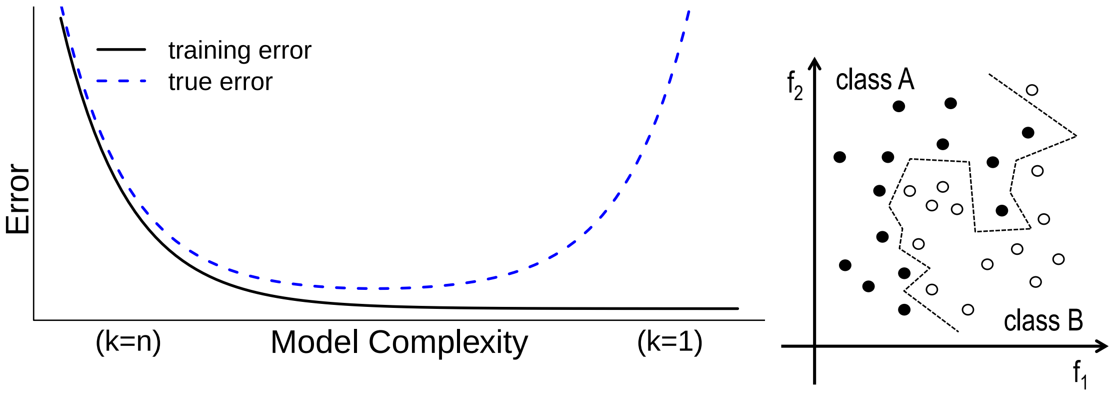
3.2.2.3 k-Nearest Neighbor (kNN)

k-NN classifier finds the k nearest points among a training set $\mathbb{S}$ and let them vote for the new data $x \in \mathbb{X}$. The k nearest neighbors are determined by computing pre-defined feature similarities (distances) between $x \in \mathbb{X}$ and ${}^{\forall}s \in \mathbb{S}$.

In k-NN, k is the factor in determining model complexity: the highest model complexity of k-NN is when $k = 1$ and the lowest model complexity is when $k = n$ [110]. For instance, if k=1, k-NN will find the most proximal data out of all the training data $\mathbb{S}$ for the classification. In the same way, for k=n, k-NN takes vote from ${}^{\forall}s \in \mathbb{S}$ data and so degenerates to the majority classifier (See Figure 3.2). With k-NN, overfitting can be minimized by choosing a larger value of k (fitting correctly all the training data of the model), while training error is minimized by choosing lower values of k. k-NN usually has good performance without a training phase but is known to be vulnerable to a bad choice of predictors [110] so it is often combined with feature selection strategies for better reliability.

3.2.2.4 Support Vector Machine (SVM)

SVM operates by treating each $s \in \mathbb{S}$ as a point in a multi-dimensional space, and then computing the hyperplanes that optimally separate the feature space into regions that are used to assign labels (See Figure 3.3). The area between hyperplanes is the '*margin*', and its width represents the generalizability of the model. A trade-off,



(a) (b)

Figure 3.2: (a) General behavior of k-NN's error terms according to the model complexity k (b) k-NN($k = 1$) tessellation graph, Example of line with $k = 1$, in 2D feature space (f_1 and f_2) for two classes (black and white dots). The highest complexity is $k = 1$ and the lowest is $k = n$ (x-axis in (a)), which is identical to majority classifier (3.2.2.1). As we see from the tessellation graph in (b), the separating surface (zigzag line) is very specialized for each training point when $k = 1$, a hallmark of overfitting.

however, exists between the margin and the error e_a: the larger the resulting interplane distance–margin, the better SVM generalizes to data outside the training data set but the training error e_a also increases as there is a rising chance of data being in-between the separating hyperplanes.

3.2.2.5 Artificial Neural Network (ANN)

Artificial neural network (ANN) is a mathematical model that attempts to simulate the structural and functional aspects of the human brain. In brain, chemical transmitters from synapses pass into dendrite raising or lowering the potential of cell. When it reaches a certain level, the cell "fires", an electrical pulse down the axon. Learning can be considered to take places in gaps between synapses. ANN attempts

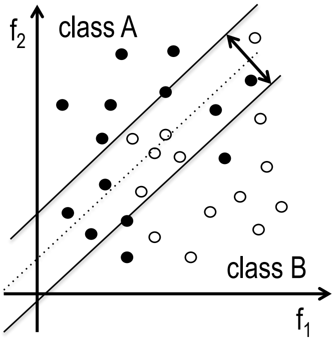


Figure 3.3: SVM Separation Plane Example. The area between the two solid lines (planes) is called the *margin*, which determines the generalizability of the model. While an increased margin gives more generalizability, it can result in less accuracy

to solve the classification problem by constructing a data structure that is similar to the biology of the human brain.

The simplest form of ANN is *perceptron*. Percentron consists of a set of node, multiple inputs and one output nodes, connected by weighted edges. As Figure 3.4 shows, an input node takes one element of feature vector f_i and is directly connected to the output node to estimate the corresponding y_i. Input feature values are multiplied by the corresponding weights and summed over the activation function for the final estimation $\hat{y}$ (Equ. 3.11).

$$\hat{y_i} = a(f_i) = \begin{cases} 1 & \text{if } \sum_{j=1 \ldots d} w_i f_{i,j} + \theta > \tau \\ 0 & \text{otherwise,} \end{cases} \tag{3.11}$$

where θ is a bias factor, τ is a pre-defined threshold level for the output node, and d is the dimension of the feature vector f_i. For the given training set $\mathcal{S} = \{(f_i, y_i) | f_i \in F, y_i \in Y\}$, the algorithm tries to find the optimal weight w_i combination to produce the closest estimate of $\hat{Y}$. The training process first initializes weights to a very small random number, then adjusts them to reduce the apparent error rate e_a by iteratively presenting the training data until convergence. With sufficient number of iterations, perceptron accomplishes a good classification on *linearly separable* problems.

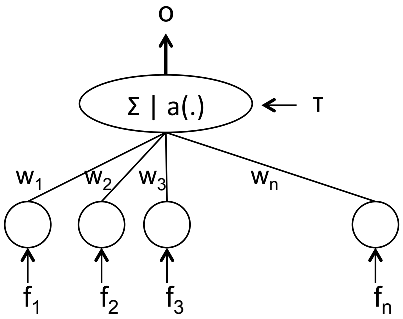


Figure 3.4: Perceptron. The simplest ANN model, perceptron, is a weighted boolean function for the given input $\mathbf{f_i} = (f_{i,1}, \cdots, f_{i,d}) \in \mathbf{F}$ and bias factor τ for the node. $a(.)$ is a activation function, which decide *firing fire*s status like a biological synapse when it is larger than the bias factor τ, a predefined threshold.

For non-linear problems, the multilayer perceptron (MLP) was developed, composed of several perceptrons in layered structure. Like the human brain, MLP connects multiple perceptrons and forms a network (See Fig 3.5). For a fully-connected feedforward network, the *backpropagation* algorithm is commonly used and detailed algorithm is well described in several textbooks [110, 135]. The goal of MLP

Algorithm 1: Perceptron Learning Algorithm.

```
Input: training data $S$
Output: Predicted label output $O$
Let $S = \{S_i = (f_i, y_i) | i = 1, \cdots, n\}$ ;
$w = w^0$;
while do
    foreach training example $S_i \in S$ do
        Compute the predicted output $o_i^k$;
        foreach weight $w_j$ do
            Update the weight, $w_j^{k+1} = w_j^k + \alpha(y_i - o_i^k) f_{ij}$;
            ( $\alpha$=learning rate);
        end
    end
end
```

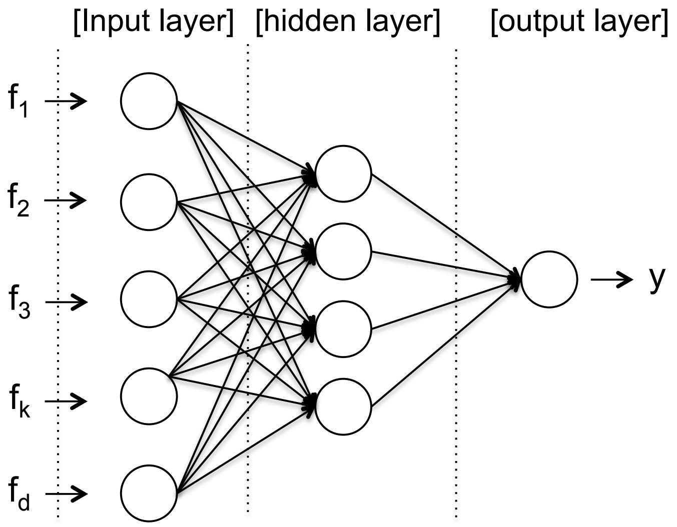


Figure 3.5: A fully connected multilayer perceptron (MLP) architecture with one hidden layer is illustrated. At the input layer, MLP has the number of input nodes equal to the number of input feature elements. Each element in the input feature is forwarded to input nodes in the hidden layer and then weighted sum propagates to the output node. MLP is originally designed for single label classification, but is easily expandable by adding the desired number of output nodes with connections to the hidden nodes.

is to estimate each weight w that minimizes the total sum of squared error (SSE):

$$\text{Error for network: } E = \sum_{k \in outputnodes} E_k$$

$$\text{Error for output unit i: } E_i = \frac{1}{2}(y_i - o_i)^2$$

Like the perceptron, once weights are initialized to very small random numbers, then the backpropagation algorithm for the *feed forward* ANN updates each weight [110] according to the following equation:

$$w_{ijk} = w_{ij(k-1)} + \Delta w_{ij(k-1)}, \ \ where \ \Delta w_{ij} = \alpha \delta_j o_i$$

where α is a network-specific learning rate. The response of the nodes is:

$$\text{output node: } \delta_j = (y_i - o_j)(o_j)(1 - o_j)$$

$$\text{hidden node} \in \mathbb{H} : \delta_j = (o_j)(1 - o_i) \sum_{k \in successors(j)} \delta_k w_{jk},$$

Detailed derivation how *backpropagation* works is also described in Appendix A.2.

By looking at the equations involved in the ANN algorithm, one can see that ANN is a lot like statistical regression with high dimensionality. In fact, each single hidden node adds a separating hyperplane, therefore ANN could result in large number of hyperplanes. There are a few design issues in ANN learning as mentioned in [110], but the most problematic issue comes from the fact that ANN is a *universal approximator* [70, 27], which can learn anything, but may not be generalizable [110]. Despite this concern and others, including overfitting, note that our previous studies [122, 84] produced very robust results for brain MR sub-cortical segmentation using single-site data.

3.2.3 Ensemble Machine Learning algorithms

Following three methods, Bagging, AdaBoost, and Random forest are all **ensemble** type methods. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples [38]. It is well-known that ensemble methods can be used for improving prediction performance. The ensemble method aims to improve accuracy of the classifier by generating a composite model of multiple classifiers, all of which are derived from the same *base classifier*. The main idea behind the ensemble methodology is to weigh several individual classifiers, and combine them in some way to obtain a new classifier that out performs the originals [137].

With the choice of a *base classifier*, ensemble method construct N classifiers,

$$C_i, \text{ for } i = 1, \cdots, N.$$

The performance of ensemble classifiers relies on the dependency between base classifiers. That is, if base classifiers are highly correlated, the prediction made out of the composite model will not improve significantly. Various strategies for constructing a set of classifiers have been proposed to produce better accuracy on the prediction [137]. In this report, we limited our investigation to the most popular resampling methodologies, *Bagging* and *Boosting*, to build high performance ensemble classifiers. A complete review of ensemble classifications are provided both in [137] and [110].

3.2.3.1 Bootstrap aggregating (Bagging)

Bagging works by taking votes from base classifiers of choice. Once the base classifier is decided, each classifier C_i is constructed on the subset $S_i \subset \mathbb{S}$, randomly resampled with replacement, i.e. *bootstrapped* from $\mathbb{S}$. Bagging then takes votes from all classifiers ${}^{\forall}C_i$ and assigns the new output. A general choice of base classifiers is a “decision stump [71]”, a tree structure classifier with one root immediately connected to the terminal node. The Bagging is known to be most beneficial with methods of high variance in their estimation, such as ANN or tree structure classifiers.

3.2.3.2 Adaptive Boosting (AdaBoost)

AdaBoost [48] works very similar to bagging but it is distinguished by building each classifier *in serial*. AdaBoost’s training proceeds by increasing the importance of sample data subset that failed in the previously built classifier. For the training data that was misclassified at run i with C_i, the importance for that data set is adaptively increased so that the successive classifier C_{i+1} can be improved. The apparent error e_a at current classifier C_i is used to weigh the data set.

Theoretically, AdaBoost can do worse if the training sample includes many misclassified points, (e.g., imperfect manual traces), since the next classifier is more adapted to the misclassification given in the training set.

Algorithm 2: Bagging Procedure [110]

```
Data: training data $S = (F, Y)$
Result: Predicted label output $\hat{Y}$
begin
    Let k be the number of bootstrap samples.;
    for $i = 1$ to $k$ do
        Create a bootstrap sample of size $N$, $D_i$ ;
        Train a base classifier $C_i$ on the bootstrap sample $D_i$ ;
    end
    $C^*(x) = \arg\max_y \sum \delta(C_i(x) = y)$;
    $\{\delta(\cdot) = 1\}$ if its argument is true and 0 otherwise;
end
```

3.2.3.3 Random Forest

Random forest [21] makes decisions from multiple tree structure classifiers. One of very appealing properties of random forest is its *generalizability*. It has been shown that the upper bound for the generalization error of random forests converges when the number of trees $\mathbb{T}$ is sufficiently large [110]:

$$\text{Generalization error} \leq \frac{\bar{\rho}(1 - t^2)}{t^2},$$

where $\bar{\rho}$ is average correlation among the trees and t is a quantity that measures the strength of the tree classifiers. More details about generalizability is summarized in Appendix A.3. As the trees become more correlated or the strength of the ensemble decreases, the generalization error bound tends to increase [110]. The random forest model is generally configured with a maximum tree depth $\mathbb{D}$, number of trees $\mathbb{T}$, and number of features to be used for split $\mathbb{F}$.

3.2.4 Related Research

A number of machine-learning techniques have been employed for the automated segmentation in the literature: SVM[129, 181, 2, 60, 72, 138, 104, 90, 178], Bagging[103], AdaBoost[129, 90, 103], kNN[32, 168, 31, 4, 33, 163], and ANN[168, 84].

SVM is a popular method investigated in multiple studies. In [181, 60, 138], SVM was employed to extract brain tumors from MR images. SVM was combined with a multiscale, multi-channel 3D segmentation algorithm in [2] to classify three different tissue types and background with regional segmentation. The method is a extension of 2D segmentation algorithm and SVM

It was one of the very early '3D' segmentation approaches that deals with 3D images as extended 2D slices instead of 3D voxels. Guo et al. [60] also presented a multiclass SVM-based segmentation tool mainly for three tissue types (white matter, gray matter, and CSF) in the brain. Segmentation of amyloid plaques in MR images of the transgenic mouse brain SVM [72]. Zhang et al. [178] employed SVM for brain MR segmentation using both grayscale (intensity) value and texture pattern [178].

Other methodologies are also found to be useful for MR brain segmentation and/or tissue classification. k-NN has been used for tissue classification with non-rigid registration [168] as well as neonatal brain segmentation [4]. De Bresser et al. [32] utilized k-NN for brain volumetric measure and in a later study [33] compared k-NN MR segmentation methods to other tools available in the field.

Random forest has only drawn researchers' attention recently, therefore there is limited research using the method. We found few studies that utilizes random forest for image processing [141, 116, 91]. One study by Yi et al. [177] presented a brain tissue classification method based on random forest.

To the best of our knowledge, there are also few comparative studies of ML algorithms for brain MR segmentation [163, 129, 33, 104].

Vaidyanathan et al. [163] compares two ML methods: k-NN and semi-supervised fuzzy c-mean (SFCM) in the context of tissue classification. Their results were unable to determine if one method was better than another in normal subjects. While the study provides a set of complete comparative discussions in the paper, the small sample size, ($n = 6$), limits the power of their conclusion. Quddus et al. [129] concluded that SVM and AdaBoost are compatible in white matter lesion segmentation, task with more time efficiency in AdaBoost. Boer et al [31] conducted a series of comparative studies regarding the tissue segmentation accuracy and reproducibility between: FMRIB's automated segmentation tool (FAST), statistical parametric mapping v.5 (SPM5), k-NN, and conventional k-NN. Their conclusion was that k-NN presents the highest accuracy but the lowest reproducibility.

While the studies above focuses more on tissue or lesion segmentation, Morra et al [104] compares four different methods for hippocampal brain MRI segmentation. They trained on 30 subjects and tested on an independent set of 40 subjects for comparison. In their study, AdaBoost and Ada-SVM outperformed both manual tracing and FreeSurfer. As they discuss in their paper, however, their validation of the segmentation algorithm is not unbiased due to intrinsic discrepancies between the training and testing data sets which may have resulted in under-estimation of the performance.

3.3 Method

In our investigation, we compute all the performance measures based on cross-validation (See Equation 3.1-3.8). For the screening study with WEKA, cross-validation was provided by WEKA where individual voxels are included/excluded. For full-scale image processing, we designed a custom cross-validation scheme where an entire subjects voxel-data are included/excluded. Subjects ($n = 35$) were roughly sub-divided into 10 subsets and cross-validation was conducted to estimate the most accurate segmentation performance. This subject-basis cross-validation provides more meaningful performance measures by allowing direct volumetric comparison.

All validity measurements of our automated segmentation results were measured against the ‘gold standard’ manual traces. Three experienced experts randomly traced six subcortical structures for each subject given T1-weighted (and T2-weighted, if applicable) images. The tracing process was blinded to clinical data such as disease status, gender, and age. The comparison of our segmentation results against to the gold standard is reported in DI, RO, HD, and ICCs (See Section 1.4 for definitions)

3.4 Experiments & Result

Results are presented in three parts: 1) screening study comparing 12 algorithm variations, 2) full-scale comparative study only between ANN and random forest, and 3) deterministic experiments for optimal region-specific normalization among 11 statistical strategies.

Experimental Data: The method described in [122, 84] is used to create the trial samples. A feature vector F is given

$$F = \{\rho_i^s, \phi_i^s, \theta_i^s, G_{i,T1}, G_{i,T2}, G_{i,SG}\}, \tag{3.12}$$

where ρ^s, ϕ^s, and θ^s are the symmetrical spherical coordinate information, $G_{i,img}$ is image intensity along the gradient descent direction of the deformed prior at the image location i [122, 84] for $img \in \{\mathcal{I}_{T1}, \mathcal{I}_{T2}, \mathcal{I}_{SG}\}$ where SG is the sum of gradient magnitude image of T1 and T2. The sample is created for each sub-cortical structure in both left and right hemispheres.

3.4.1 Screening Study: Machine-Learning Comparison with WEKA

Experimental Set-Up: 12 algorithm variations of machine-learning were contrasted on the MR segmentation experimental data (Equ. 3.12). Four sub-cortical structures including caudate, hippocampus, thalamus, and putamen, are examined. For all eight algorithms, we employ WEKA, a publicly available machine-learning tool [62], for efficient comparison which exploits the identical MR segmentation data for fair comparison. Table 3.2 displays one of the representative results for the caudate.

The WEKA-provided default first initiates the experiments. Further variations are then expanded progressively if one exhibits improvement over the initial performance. Favored results with the default setting led us to identify K-NN, ANN, and random forest as suitable method (Table 3.2). Each extended experimental design grounds on either theoretical, previous [84], or pre-conducted empirical study. Including the extended experiments, results of the 12 machine-learning approaches are summarized in Table 3.2.

Results: The screening study identified four favorable algorithms in sub-cortical segmentation: Bagging, k-NN, ANN, and Random forest (Table 3.2). The assets of four algorithm are well supported by five metrics reported in the WEKA (Table 3.2) and results of other structures including hippocampus, putamen, and thalamus are presented in Appendix Table A.2, Table A.3, and Table A.4 as well. Table 3.2 shows one of results from caudate for each in left (L) and right (R) hemisphere, background (Bg), and averaged performance(Avg). Note that other than k-NN, all three favored methods belong to Ensemble category.

Those favored methods are decided based on the metrics of sensitivity, specificity, precision based on confusion matrix (See Section 3.2.1.1) as well as additional ones, F-Measure and AUC, that WEKA reports:

- F-Measure $= \dfrac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})}$
- AUC = Area under the curve

3.4.2 Full-Scale ANN and Random Forest Comparative Study with OpenCV

Candidates Selection: For the full-scale investigation of machine-learning algorithms, we narrow down to ANN and random forest. From the screening study in Section 3.4.1, we identified four favorable techniques: Bagging, k-NN, ANN, and random forest. Of four methods, we only include random forest and ANN based theoretical and empirical studies as following.

ANN and random forest are favored by several reasons. First, both ANN and random forest displayed excellent performance than others regardless of their parameter variation (Table 3.2). Second, ANN has already proved its high accuracy for sub-cortical segmentation in the previous research [84]. Third, two preferred properties of random forest, convergence and generalizability [20], are attractive for further investigation. Those two properties are attractive especially for the research involving large variation in MR data. Upon those reasons, ANN and random forest are favored for additional investigation.

Table 3.2: Screening study contrasting performance of 12 machine learning approaches for the identical MR segmentation data of the caudate in WEKA.

Location	Sensitivity	Specificity	Precision	F-Measure	AUC	Sensitivity	Specificity	Precision	F-Measure	AUC
			Majority Classifier					**Bagging**		
L	0	0	0	0	0.500	0.876	0.020	0.881	0.879	* **0.991**
R	0	0	0	0	0.500	0.880	0.021	0.887	0.883	* **0.991**
Bg	1	1	0.702	0.825	0.500	* **0.951**	0.122	* **0.948**	* **0.950**	* **0.979**
Avg	0.702	0.702	0.492	0.578	0.500	* **0.929**	0.092	* **0.929**	* **0.929**	* **0.982**
			Naïve bayes					**kNN (k=1)**		
L	0.298	0.1	0.334	0.315	0.821	0.882	0.027	0.846	0.864	* **0.930**
R	0.717	0.272	0.325	0.447	0.807	0.894	0.028	0.852	0.872	* **0.933**
Bg	0.720	0.086	* **0.951**	0.819	* **0.901**	* **0.933**	0.112	* **0.951**	* **0.942**	* **0.911**
Avg	0.659	0.117	0.766	0.689	0.875	* **0.919**	0.087	* **0.921**	* **0.920**	* **0.917**
			SVM					**kNN (k=10)**		
L	0.776	0.044	0.749	0.762	0.866	* **0.925**	0.030	0.836	0.878	* **0.990**
R	0.777	0.051	0.735	0.755	0.863	* **0.902**	0.023	0.877	0.889	* **0.991**
Bg	0.885	0.224	* **0.903**	0.894	0.831	* **0.935**	0.087	* **0.962**	* **0.948**	* **0.980**
Avg	0.853	0.171	0.855	0.854	0.841	* **0.928**	0.069	* **0.931**	* **0.929**	* **0.983**
			AdaBoost					**kNN (k=20)**		
L	0	0	0	0	0.203	* **0.902**	0.027	0.848	0.874	* **0.991**
R	0	0	0	0	0.791	* **0.919**	0.031	0.845	0.881	* **0.992**
Bg	1	1	0.702	0.825	0.507	* **0.930**	0.089	* **0.961**	* **0.945**	* **0.980**
Avg	0.702	0.702	0.492	0.578	0.507	* **0.924**	0.071	* **0.927**	* **0.925**	* **0.983**
			ANN (HN=20)					**Random forest** ($NT = 10$)		
L	0.880	0.024	0.860	0.870	* **0.987**	0.898	0.023	0.866	0.882	* **0.987**
R	* **0.901**	0.030	0.848	0.873	* **0.985**	0.864	0.016	* **0.906**	0.884	* **0.988**
Bg	* **0.935**	0.109	* **0.953**	* **0.944**	* **0.969**	* **0.952**	0.119	* **0.949**	* **0.950**	* **0.976**
Avg	* **0.922**	0.085	* **0.923**	* **0.922**	* **0.975**	* **0.930**	0.090	* **0.931**	* **0.930**	* **0.979**
			ANN (NH=60)					**Random forest** ($NT = 25$)		
L	0.891	0.020	0.885	0.888	* **0.989**	0.884	0.018	0.894	0.889	* **0.991**
R	* **0.904**	0.022	0.881	0.893	* **0.988**	0.892	0.019	0.897	0.895	* **0.992**
Bg	* **0.949**	0.102	* **0.956**	* **0.953**	* **0.976**	* **0.956**	0.112	* **0.953**	* **0.954**	* **0.982**
Avg	* **0.934**	0.078	* **0.934**	* **0.934**	* **0.980**	* **0.936**	0.084	* **0.936**	* **0.936**	* **0.985**

Five measurements are reported for left (L), right (R), background (Bg) of caudate, and average of all three regions (Avg): sensitivity, specificity, precision, F-measure, and area under the curve (AUC). * indicates a metric > 0.9 (the desired performance/success rate).

Bagging and k-NN are excluded for their undesired properties despite of their relatively higher accuracy in the screening study (Section. 3.4.1). k-NN is refrained from more investigation because of its susceptibility to the noise, or data variation. This unreliable feature is particularly dangerous for scalable data processing by limiting model's generalizability. Next, bagging is excluded to avoid duplicative effort to random forest. That is, the basic setting of bagging in WEKA utilizes a decision stump as a base classifier. Since the decision stump is one-depth decision tree, we concluded that it could create extra effort. For those reasons, both k-NN and bagging are leaved out of subsequent consideration.

Experimental Set-Up: Cross-validation on a subject-basis is conducted for full-scale comparison study. The results is reported in Figure 3.7 for caudate, putamen, thalamus, and hippocampus. The parameter variation includes two ANN's hidden nodes $\mathbf{H} = 20$ and 60 and three random forest's number of trees $\mathbf{T} = 10, 25,$ and 100, with the depth of tree $\mathbf{D} = 100$. The choice of parameter for ANN is based on the identical reasoning from the screening study. To choose the parameters of the random forest $\mathbf{T}$ and $\mathbf{D}$, however, we also conduct exhausted parameter search as shown in Figure 3.6. From the experiments, we conclude that the depth of trees could be large enough $\mathbf{D} = 100$, while $\mathbf{T}$ has to be tested more carefully, including three variations as mentioned above. Five correspondence measures between automated and manual delineations for all four regions of interest in left hemisphere are reported. The measurements includes relative overlap (RO), dice index (DI), Hausdorff distance (HD), and intraclass correlation of agreement ($ICC(A)$) and consistency ($ICC(C)$) (Definitions in Section 3.3). In addition to those five measures, asymmetry index between left and right hemisphere is also reported to highlight performance consistency.

$$AI = \frac{|A - B|}{|A + B|} \tag{3.13}$$

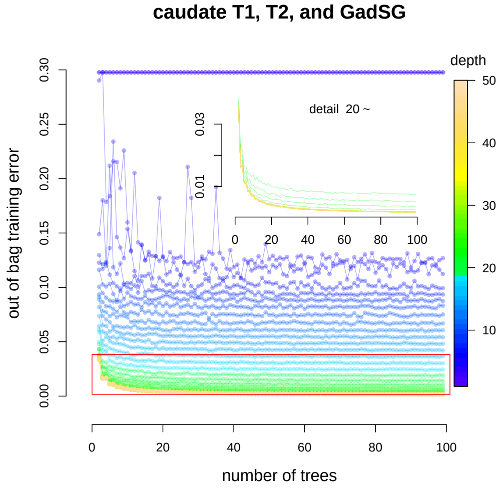


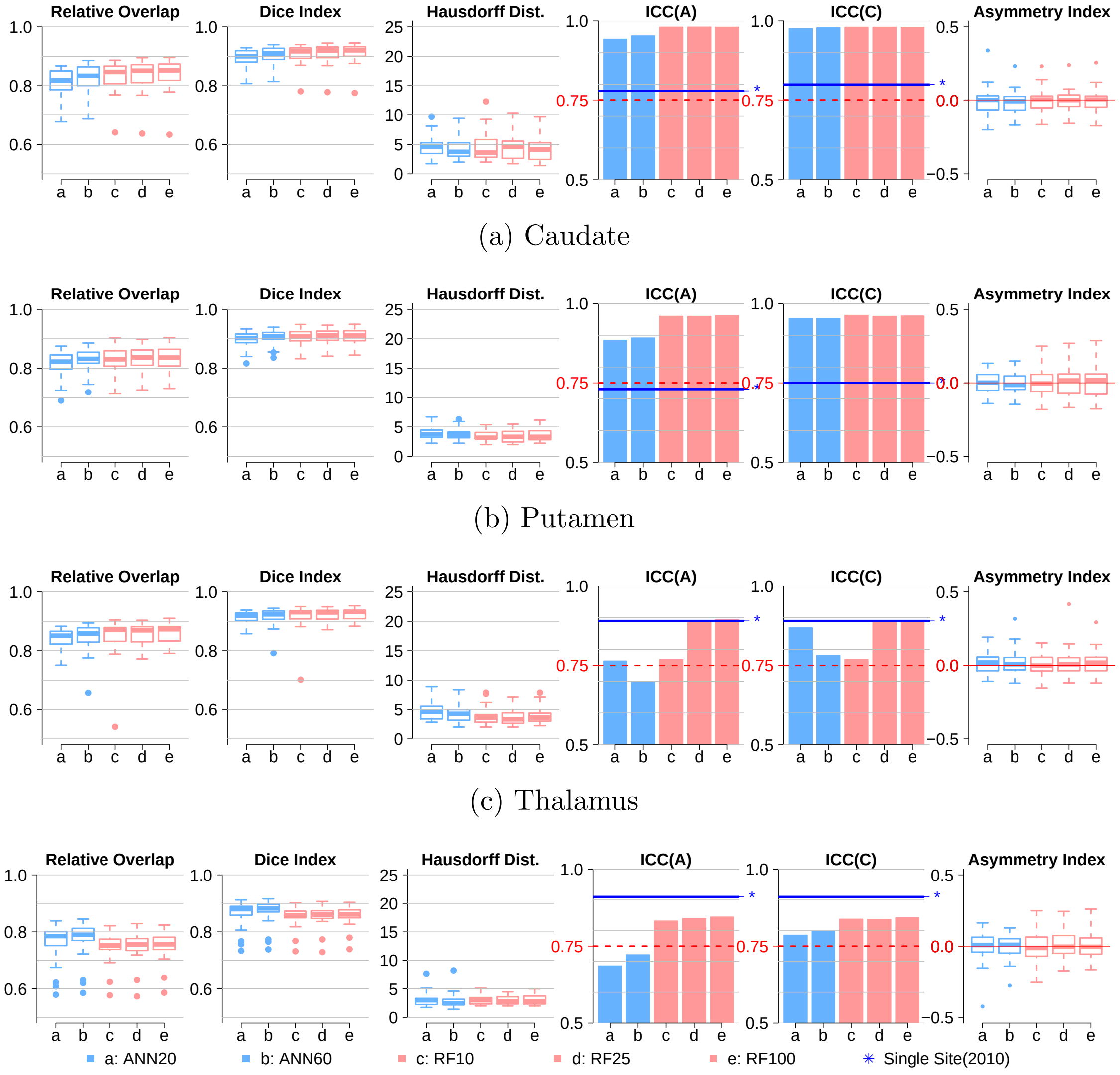
Figure 3.6: Out of bag error computation for left and right caudate with T1, T2, and F_{SG}. Convergence behavior is observed along the number of trees and depth of trees. The upper right graph shows an enlarged graph for the number of depth $\mathbf{D} > 20$.

Results: In general, the performance is more discernible with both $ICCs$ than other three measures: RO, DI, and HD. Even though the performance is indistinguishable in RO, DI, and HD, all three measures present high correspondence between manual and automated method. Regarding both ICC values, a series of random forest variation is preferred to ANN in all four regions of interest.

It is hard to differentiate performance by RO, DI, or HD (three measures from the left), but ICC measures generally suggest that random forest is superior. With left caudate and putamen (See Figure 3.7a 3.7b), the both machine-learning methods are well above both 0.75 (red dashed line) and single-sited study [84] (blue solid line). For left thalamus (Figure 3.7c), however, we see that only random forest of $\mathbf{T} = 25$ and 100 exhibits as much accuracy as single-sited performance (solid blue lines with $\star$). Last, hippocampus (Figure 3.7d) presents superior performance with regard to ICCs. Even though random forest outperformes ANN in hippocampus trial, ICC measures are far below the single sited study. Note that, however, random forest performance on hippocampus still above 0.75. In conclusion, the results poses strong support to random forest against ANN in processing multi-site data. Consistent asymmetry index for both structures also could be highlighted in that right structure segmentation result is well corresponding to the left structure's one presented here. Dashed red line at 0.75 in both ICC plots represents a bottom line suggested by Shrout et al. [145], where two independent traces, manual and automated one, can be regarded as identical. Solid blue line with a star($\star$) mark showed the previous study result [84], which highly optimized for **single-sited study**.

3.5 Discussion

Random forest Over ANN: We have selected the random forest algorithm over ANN. Both the ANN and the random forest algorithm achieved sufficient high segmentation accuracy, which is measured in correspondence to the manual traces via 10-folds cross validation. Even though they presented compatible results, there



(a) Caudate

(b) Putamen

(c) Thalamus

(d) Hippocampus

Figure 3.7: Five correspondence measures between automated and manual delineations for four structures in left hemisphere with asymmetry index between left and right structures in the far right side. Five measures includes relative overlap (RO), dice index (DI), Hausdorff distance (HD), and intraclass correlation of agreement ($ICC(A)$) and consistency ($ICC(C)$). Left two measures (light blue) in each graph are ANN trials with $\mathbf{H} = 20$ ($ANN20$) and $\mathbf{H} = 60$ ($ANN60$) and right most three meausures (light red) are random forest trial with $\mathbf{T} = 10$ ($RF10$), $\mathbf{T} = 20$ ($RF25$), and $\mathbf{T} = 100$ ($RF100$). Dashed red line at 0.75 is suggested minimum by Shrout et al. [145] and solid blue line showed the previous study result [84].

are two primary properties that we favored the random forest algorithm to ANN: 1) the overfitting possibility of ANN and 2) the great generalizability of random forest. These properties, the overfitting and the generalizability of ANN and random forest, respectively, are contrary each other and theoretically well supported. Generalizability of the algorithm is desired for large-scale multicenter data processing while the overfitting is ill-favored. These properties are evidently observed from our experiments. For instance, the overfitting phenomena generally resulted in performance declining along with the increased model complexity and it is clearly observed from the experiment of ANN presented in Figure 3.7. For the thalamus in Figure 3.7, the segmentation accuracy falls back as the ANN's model complexity goes up. On the other hand, the random forest algorithm converges toward higher segmentation accuracy with the increased number of trees. Theoretical properties of each ML algorithm were confirmed via our comprehensive experiments and lead us to choose the random forest algorithm over ANN.

Limitations:

One disadvantage of random forest is that the size of the model grows quite large compared to the other models. The trained model size for random forest is approximately 2GB for $\mathbb{H} = 100$, and that can overload hardware resources if they needed to run in a low memory processing environment (Figure 3.8).

The comparison study is obviously limited, as there are many other algorithms and parameters that could be evaluate in a similar manner. We, however, restrict our initial investigation for the classification methods generally used in the field and even

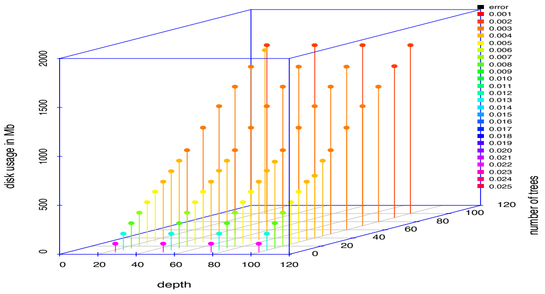


Figure 3.8: Memory Usage according to the number of trees $\mathbb{T}$ and the depth of the tree $\mathbb{D}$. The disk usage is more depend on the $\mathbb{T}$ and increase up to 2GB.

further restrict based on the preliminary investigations in WEKA before advancing the most likely candidates to full-scale implementation and investigation.

While literatures in the field commonly shows good performance of SVM [181, 2, 60, 122, 138, 104], SVM was excluded from further full-scale investigation due to its poor default performance (Section 3.4.1). In addition to the poor default performance, because of time and resource restriction, SVM has not been investigated in depth in this study.

3.5.1 Summary & Conclusion

An automated segmentation framework of six subcortical structures from MRI is developed by investigating ML algorithms. The exceptional segmentation accuracy and generalizability were accomplished with the random forest algorithm comparing to other 10 we tested. The segmentation framework was tested with the series of experiments to compare 11 ML algorithms. To make test effective, 11 ML algorithms are first screened by the publicly available ML tool, WEKA, with the down-sampled in-vivo MR segmentation data. Extensive comparative studies between ANN and the random forest are then performed and revealed the superiority of the random forest algorithm. Even though the framework is highly tuned for the MRI subcortical segmentation, it can easily be adapted to investigate ML algorithms for segmentation work in other domains as well. The developed framework was successfully applied on two independent on-going large-scale clinical studies [127, 161] that includes more than 3000 scans.

CHAPTER 4
OPTIMAL INTENSITY NORMALIZATION SELECTION

Intensity normalization is commonly practiced but has not been explored extensively for medical image processing. Here we review commonly used intensity normalization functions and propose a region-specific localized normalization from a choice of the most promising normalization function. 11 variations of intensity normalization functions are compared and analyzed for their contribution toward accurate segmentation of six subcortical structures. All the normalization functions generally enhanced the segmentation accuracy. Among the 11 cases, the double-sigmoid function with parameters set by 5^{th} and 95^{th} quantiles resulted in superior segmentation accuracy. We also observed that depending on the relative location of regions of interest, the intensity profile varies which, in turn, resulted in performance variation according to the choice of normalization. The performance dependency on the normalization functions is discussed with regard to intensity profiles of each region of interest.

4.1 Introduction

Normalization is a commonly practiced and required preprocessing step for robustness of subsequent phases in many data processing technologies. In statistics, normalization is formally defined as a standardization of data obtained from different sources at different periods, through comparison against the objectives of data collection. Paradigm of normalization is introduced to address difficulties in data processing specific to image data variation. The objective is to reduce variability by transforming measured intensities to a common compatible scale without altering clinically relevant information. This step is leveraged to improve robustness of the succeeding stages by eliminating serious inhomogeneity within the data.

Normalization in image processing, especially with the growing interest in multi-site and large-scale study design in recent years, should benefit ultimate outcomes by regulating data variation at the beginning of the analysis. Instrument and experimental influences can bring systematic and random variations in the data collection results (signal intensity of MRI data set). Commonly observed contributers are varying manufacturers, field strengths, and MR acquisition protocols which alter MR image intensity profiles. It is possible to minimize those data variation with careful designing but impossible to entirely remove their effect due to the inherent limitations of MRI. Since multi-site/large-scale study potentially introduces greater data variation, it is important to take guarded consideration into account.

It is accepted that normalization is a fundamental preprocessing step for robust image processing, however there has been little or no studies done investigating the impact of different normalization methods with respect to robust statistics. The normalization techniques applied to MRI processing are often limited to basic and routinely used procedures in the field such as histogram equalization or traditional statistical techniques including standardization and/or linear scaling. Moreover, these normalization techniques either introduce a bias to a selected template (in the case of histogram equalization) or rely on a set of assumptions that may not be fulfilled, including linearity, normality, and independence (in the case of standardization). Several studies warn of the danger of ignoring those strong assumptions in the statistical procedure [65, 44]. We observed a reasonable amount of unexpected behavior in the algorithm mainly due to data variation. For all these reasons, it is highly doubtful that those commonly practiced normalization technologies are best suited for multi-center, large-scale MR data processing.

To this end, we investigated the influence of various normalization procedures on the segmentation framework. Specifically, we intended to quantify the influence of region-specific robust statistics for the normalization methods in a large-scale MR image processing setting. Different normalization techniques were compared and the underlying assumptions are discussed in this report.

4.2 Background

We aim to find an adequate normalization method together with robust statistics that are best suited for the brain MRI subcortical segmentation. To begin, we introduce and discuss related background: robust statistics, normalization methods that commonly used in image processing techniques, and two normalization paradigms: global versus local.

4.2.1 Robust Statistics

Robust statistics aims to provide an accurate description of the data in the presence of gross error. Hampel [64] asserts that robust statistics are the stability theory of statistical procedures. Robustness can be quantified via the concept of the *"breakdown point."* As defined by Hampel, the breakdown point describes the largest fraction of arbitrary gross errors tolerated before the statistic *"breaks down"* and becomes totally unreliable. Typically, breakdown point is a function of sample size n, but the asymptotic breakdown point when n is an arbitrary large number is commonly calculated for assessment [53].

One conventional example that illustrates the necessity of robust statistics is taking the difference between mean and median. Median is regarded as more resistant to a gross error than mean: median provides a robust description of the data with up to 50% gross error whereas that for the mean is 0% [136]. Mean and median are examples of a *locational* or so-called *central tendency estimator* for data with different degrees of tolerance to outliers. Similarly, there are statistics more robust for data dispersion than standard deviation and range ($maximum - minimum$), such as interquartile range (IQR), and median absolute deviation (MAD). Those measures are examples of a *dispersion parameter* or *dispersion estimator*.

We will focus on robust statistics of locational and dispersion measurements for a robust MR intensity normalization. There are several methods other than those introduced in this chapter, that are designed for robust estimation, such as the expectation-maximization (EM) algorithm. In the present study, however, we limited our investigations to the robust statistics that are reasonably efficient to compute, such that the normalization procedure can be integrated into our segmentation framework in reasonable time. In the following, we will discuss the assumptions and characteristics of each statistic in the context of MR image segmentation.

4.2.1.1 Location parameters Θ

Location parameter is a measure of the central tendency for a given distribution. Most commonly used location parameters are *mean* and *median*:

Sample mean $\bar{x}$: Sample mean is the most commonly used one in the field however the value is sensitive to outlier, which has 0% breakdown point.

$$\bar{x} = \frac{1}{n}\sum_{\forall x} x$$

Sample median $Q_{n/2}$: Sample median is the most common alternative to mean. With 50% breakdown point, it is less sensitive to outliers than mean $\bar{x}$.

4.2.1.2 Dispersion parameters Φ

A dispersion parameter, or scale parameter, describes the dispersion or scale of the distribution:

Range R: Range is the simplest dispersion descriptor which is calculated as the difference between minimum and maximum of the samples:

$$R = max(x) - min(y)$$

Standard deviation s^2: Standard deviation is one of the most used dispersion parameters:

$$s^2 = \frac{1}{n}\sum(x - \bar{x})^2$$

Interquartile range IQR: While range R and sample variance s^2 have breakdown point of 0%, the interquartile range has 25%:

$$IQR = Q_{3n/4} - Q_{n/4}$$

4.2.1.3 Modern robust statistics

Trimmed mean and variance: Trimmed mean is the mean of the central $n \cdot \alpha$ part of distribution, so $n \cdot \alpha$ observations are removed from each end. Trimmed mean is alternative to mean or median as a compromised estimator between the mean and median [44]

Winsorized mean and variance: While trimming ignores values outside a certain range, winsorizing brings in extreme observations of $n \cdot \alpha$ to some constant. The benefit of the winsorized variance is that it is more resistant to outliers than variance is and can result in more accurate standard errors than classical methods [44]

4.2.2 Region-specific Normalization based on Robust Statistics

Region-specific normalization is devised to maximize information consistency across scans from various sources (sites and/or scanners) while enhancing image contrast for better separation of neuroanatomy. The proposed region-specific normalization employs both robust statistics and spatial localization via deformed subject-specific priors (Sec. 1.6.1). We describe the theory of region-specific normalization (Sec. 4.2.2.1) and review both robust statistics and normalization methods investigated in this study (Sec. 4.2.2.2).

4.2.2.1 Region-specific Normalization

Region-specific normalization is designed to enhance structural details of an MR image $\mathcal{I}$ for a focused region $\mathcal{R} \subset \mathcal{I}$. For each label $l \in \mathbb{L}$, the focused region is identified by the subject-specific prior p_l generated by deforming the template spatial prior with a high-deformable registration $\mathcal{T}$ into subject space $p_l(\mathcal{T}(x))$.

The method takes into account all of the *locally computed statistics* within the spatially bounded region $\mathcal{R}_l$:

$$\mathcal{R}_l = \{x | 0 < p_l(\mathcal{T}(x)) < 1, x \in \mathcal{I}\}. \tag{4.1}$$

Note that most normalization techniques in MRI processing utilize statistics computed globally to deal with intra-scan intensity variations, e.g., histogram equalization between two images. This *global normalization* method, however, is less sensitive to specific regions of interest. In this study, we hypothesize that normalization methods with region-specific statistics would enhance the robustness of the segmentation framework in our multi-site longitudinal data processing environment where large intensity variations are expected. In other words, much more anatomical details in MR image can be enhanced by region-specific normalization and provide increased consistency across scans. A 3D example of warped region-specific priors for the caudate nucleus in both left and right hemispheres is shown in Figure 1.8.

4.2.2.2 Normalization with Robust Statistics

11 normalizations including parameter variations were investigated and evaluated in this paper. For each normalization function $\mathcal{N}$, the robust statistics (Sec. 4.2.1.3) aim to provide an accurate data description in the presence of gross error. Normalization equations with those robust statistics are summarized in Table 4.1. Note that normalization functions used in this study are based on either simple linear scal-

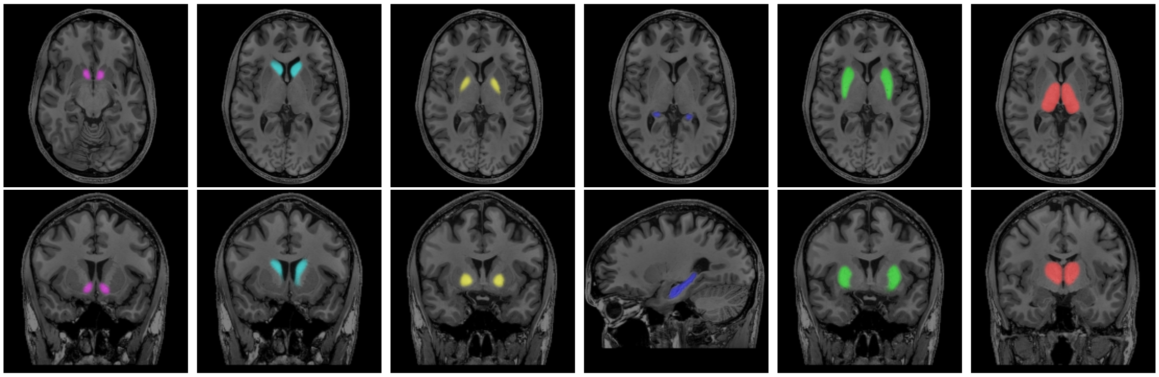


Figure 4.1: Warped spatial priors (colored) on top of a subject (gray scale) with the explanatory figure for localized intensity normalization. 3D surfaces (purple) were generated for both region A and B to show size of search area. The yellow to red colors represents spatial prior value from 0.01 to 1.00. Image intensity is linearly transformed based on locally identified minimum and maximum value in the BRAINSCut feature generation process. Region A and B have different linear scaling parameters where computed independently for each region.

ing or sigmoidal transformation as shown in Figure 4.2. A region-specific localized normalization method was devised to enhance intensity characteristics for specific regions of interest. More region-specific details can be captured by locally restricting normalization with robust statistics.

Five groups of normalization functions $\mathcal{N}$ (linear scaling, sigmoid, double sigmoid, trimming, and winsorizing) were employed for the comparative experiment. Table 4.1 displays all five equations with necessary parameters and Figure 4.2 shows the major shapes of the normalization functions. Note that with regard to the transformation shape linear scaling, trimming, and winsorizing share the same shape (Figure 4.2a).

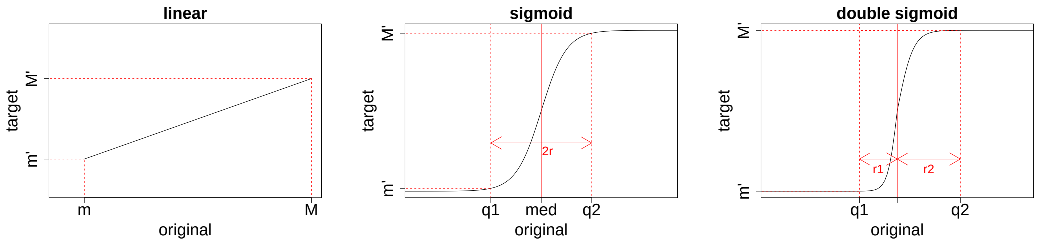
Linear scaling is the most commonly practiced standardization mechanism

Table 4.1: Five normalization functions

Method	Transform function $\mathcal{N} : \mathbb{R} \to \mathbb{R}\ \mathcal{N}(x) = x'$
Linear	$x' = a + \frac{x-b}{c}d$
α-Sigmoid	$x' = \frac{1}{1 + exp(-8 \cdot \frac{x - q_{1/2}}{r})},$
α-Double Sigmoid	$x' = \frac{1}{1 + exp(-8 \cdot \frac{x - q_{1/2}}{r})}, r = \begin{cases} r = r_1, \text{ if } x < q_{1/2} \\ r = r_2, \text{ otherwise} \end{cases}$
α-Trimming	$x' = \frac{x - \bar{x}_{t(\alpha)}}{s_{t(\alpha)}},$ $\begin{cases} \bar{x}_{t(\alpha)} = & \frac{1}{\sum \mathbb{1}_i} \sum \mathbb{1}_i x_i \\ s_{t(\alpha)} = & \frac{1}{\sum \mathbb{1}_i} \sum \mathbb{1}_i (x_i - \bar{x}_{t(\alpha)})^2 \end{cases} \quad \mathbb{1}_i = \begin{cases} 0 & \text{if } q_\alpha < x_i < q_{1-\alpha} \\ 1 & \text{otherwise} \end{cases}$
α-Winsorizing	$x' = \frac{x - \bar{x}_{w(\alpha)}}{s_{w(\alpha)}},$ $\bar{x}_{w(\alpha)} = \frac{1}{\sum \mathbb{1}_i} \sum (\mathbb{1}_i x_i + \mathbb{1}_i^l c^l + \mathbb{1}_i^u c^u)$ $s_{w(\alpha)} = \frac{1}{\sum \mathbb{1}_i} \sum (\mathbb{1}_i (x_i - \bar{x}_{w(\alpha)})^2 + \mathbb{1}_i^l (c^l - \bar{x}_{w(\alpha)})^2 + \mathbb{1}_i^u (c^u - \bar{x}_{w(\alpha)})^2)$ $\mathbb{1}_i^u = \begin{cases} 0 & \text{if } x_i < q_\alpha \\ 1 & \text{otherwise} \end{cases}, \mathbb{1}_i^l = \begin{cases} 0 & \text{if } x_i > q_\alpha \\ 1 & \text{otherwise} \end{cases}$ $c^{l,u}$ is upper/lower constant for winsorizing

Five normalization functions $\mathcal{N} : \mathbb{R} \to \mathbb{R}$ given by $\mathcal{N}(x) = x'$ for robust transformation of training data into proper scale. All five normalization functions with a few parameter variations are contrasted in this comparative study to identify the robust statistical normalization procedure for the large-scale multicenter segmentation framework.

that rescales original values in (m, M) to (m', M'). For its computational simplicity, linear scaling is very sensitive to outliers presented in the data. To compensate, trimming and winsorizing methods can provide alternatives to the original linear transformation. Both methods, however, create discontinuity in the transformed data set, which is usually an undesired property in machine-learning. On the other hand, the sigmoid and double sigmoid functions, which resemble the shape of the letter S, produce data that is continuous while dealing with outliers better. The double sigmoid function is more suited for skewed data while the sigmoid is better suited for evenly distributed data.



(a) (b) (c)

Figure 4.2: Major shapes of normalization functions used in this comparative study. While linear scaling (Figure 4.2a) provides a computationally simple and intuitive transformation, it often suffers when outliers present in the data set. There are multiple different method that uses linear scaling: simple linear scaling, trimming, and winzorizing. Note that trimming and winzorizing can result in discontinuous data. On the other hand, sigmoid type functions (Figure 4.2b and (Figure 4.2c), transform input data with less sensitivity to outliers while maintaining continuity. All of these functions are designed for comparative experiments to identify the robust normalization function for our segmentation framework. Detailed parameters for experiments are presented in Table 4.2

4.3 Experiment and Result

4.3.1 Experimental Setup

A series of experiments are designed to analyze effect of 11 normalization approaches with in-vivo MRI with regard to the subcortical segmentation framework. In this experiment, 10-fold cross-validation is used for unbiased assessment of both benefits and downside of 11 normalization strategies on each ROI. The 35 scans that are used in Section 3 are again utilized in this experiments For rigorous evaluation and modeling of the intensity normalization approaches, we have utilized all the equations noted in Table 4.1 into the experiments and compared each contribution toward six subcortical segmentation. The experimental set-up is summarized in Tabel 4.2.

4.3.2 Results

Comparative study results from 10-fold cross validation of 11 normalization strategies as well as one without normalization are contrasted in Figure 4.3. For 35 testing scan sessions, T1- and T2-weighted images and summed image of gradient magnitudes from both are used in all experiments. We have computed multiple comparative metrics including relative overlap, dice index, Hausdorff distance, average Hausdorff distance, and intraclass correlation of agreement and consistency, here we only reports ICCs, which convey the most representative information.

The experimental results are summarized in Figure 4.3. The two ICCs (See Section 1.4) are reported in the result that ordered by average ICCs across six subcortical structures per normalization function. The larger $ICC(A)$ means more agreeable

Table 4.2: Comparative study experimental setup for normalization functions

abbreviation	*function*	*parameters*
Min/Max	linear	$a =$ minimum, $b =$ maximum, $c = R$, and $d = R'$
Z-Score	linear	$a = 0,\ b = \bar{x},\ c = s,\ d = 1$
IQR	linear	$a = 0,\ b = Q_{1/2},\ c = IQR,\ d = 1$
01-sigmoid	α-sigmoid	$r = (Q_{99^{th}} - Q_{01^{th}})/2$
05-sigmoid	α-sigmoid	$r = (Q_{95^{th}} - Q_{05^{th}})/2$
01-doubleSigmoid	α-double sigmoid	$r_1 = (Q_{1/2} - Q_{01^{th}}),\ \text{if } x < Q_{1/2}$ $r_2 = (Q_{99^{th}} - Q_{1/2}),\ \text{otherwise}$
05-doubleSigmoid	α-double sigmoid	$r_1 = (Q_{1/2} - Q_{05^{th}}),\ \text{if } x < Q_{1/2}$ $r_2 = (Q_{95^{th}} - Q_{1/2}),\ \text{otherwise}$
01-Trimming	α-Trimming	$\alpha = 1$
05-Trimming	α-Trimming	$\alpha = 5$
01-Winsorizing	α-Winsorizing	$\alpha = 1$
05-Winsorizing	α-Winsorizing	$\alpha = 5$

Normalization function comparative study experimental setup for parameterization of functions given in Table 4.1. 11 normalization variation with five transformation functions are devised for this experiments to find the robust one for our segmentation framework processing large-scale multicenter MR data. Briefly, mean, median, trimmed mean, and winsorized mean are tested for locational parameter and range, IQR, standard deviation, $(Q_{99^{th}} - Q_{01^{th}})$, and $(Q_{95^{th}} - Q_{05^{th}})$ are tried for dispersion parameter.

absolute volume to the gold standard, and the larger $ICC(C)$ reflects more consistent relative relation with the gold standard, where an additive transformation serves to equate them [97]. In both cases, the larger ICC is, the superior the segmentation accuracy is in terms of the correspondence to the gold standard. The result is organized by the performance in decreasing order from top to bottom.

11 normalization function variations are all served to promote accurate subcortical segmentation. Among those 11 approaches, two normalizations of using *'trimming'* and *'linear (min/max)'* display inferior to others. Note that double sigmoid transformation with $\alpha = 01$ presented the best correspondence of the automatic segmentation to the manual trace. Although *double sigmoid* transformation with $\alpha = 01$ and 05 are ranked at the highest performance (Figure 4.3), note that the other eight normalization $\mathcal{N}$ functions show vary similar performance each other in improving segmentation accuracy.

The improvement upon the region-specific normalization technique is formally tested by a two-tailed paired t-test. Relative overlaps between manual and auto delineation from each approach from the 10-cross validation experiment are paired and statistically tested against the one without normalization. The result indicated that there are statistically significant differences in relative overlap between with and without normalization for subcortical segmentation, as p-values shown in Table 4.3. All the improvements with 11 normalization methods are statistically significant and the smallest p-value are observed with the top ranked approach, *Double Sigmoid*. In other words, the region-specific normalization approach with any choice from 11

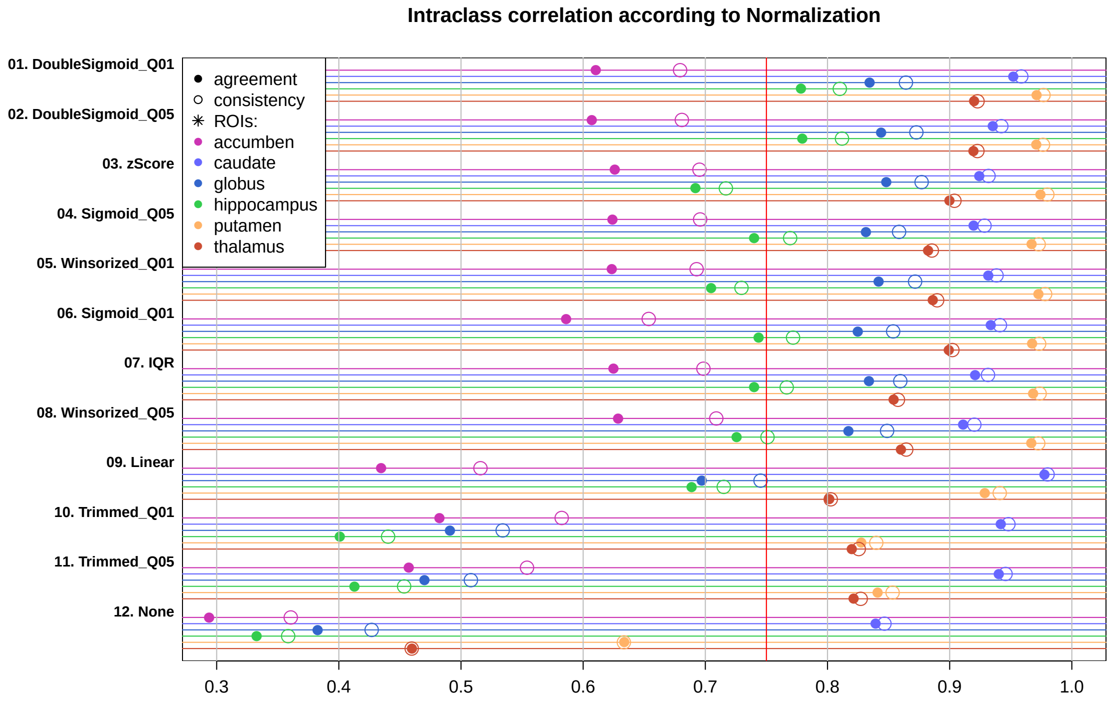


Figure 4.3: $ICC(A)$ (solid circle) and $ICC(C)$ (empty circle) dot graph is shown for 11 normalization strategies as well as raw data without normalization (None). All six structures are tested and plotted with different colors. ICCs lower bound suggested by Shrout [145] also presented as a red line. 12 methods' name on the left-hand side are ranked by its average performance over six structures from the top. That is, *'01 and 02. double sigmoid'* and *'03. ZScore'* presented top three best average performance over six structures based on ICC. Also note that any normalization benefits the performance than the one without normalization *'12. None'*.

functions effectively advances segmentation accuracy and also the rank of ICC is well reflected in the statistical test.

Table 4.3: All 11 normalization strategy significantly improved segmentation accuracy in respect to the relative overlap to the manual traces.

method	*accu*	*caud*	*glob*	*hipp*	*puta*	*thal*
D. Sig (Q^{01th})	$1.57\cdot10^{-4}$	$3.42\cdot10^{-3}$	$2.77\cdot10^{-5}$	$1.03\cdot10^{-3}$	$2.26\cdot10^{-5}$	$1.62\cdot10^{-4}$
D. Sig (Q^{05th})	$1.94\cdot10^{-4}$	$6.52\cdot10^{-3}$	$1.17\cdot10^{-5}$	$9.67\cdot10^{-4}$	$1.67\cdot10^{-5}$	$1.46\cdot10^{-4}$
zScore	$2.80\cdot10^{-4}$	$8.73\cdot10^{-3}$	$1.82\cdot10^{-5}$	$1.64\cdot10^{-3}$	$1.17\cdot10^{-5}$	$8.24\cdot10^{-4}$
Sig. (Q^{05th})	$1.62\cdot10^{-4}$	$1.31\cdot10^{-1}$	$1.61\cdot10^{-5}$	$1.11\cdot10^{-3}$	$1.73\cdot10^{-5}$	$3.60\cdot10^{-4}$
Winsor. (Q^{01th})	$7.79\cdot10^{-4}$	$5.24\cdot10^{-3}$	$1.07\cdot10^{-5}$	$1.40\cdot10^{-3}$	$1.40\cdot10^{-5}$	$1.21\cdot10^{-3}$
Sig. (Q^{01th})	$1.38\cdot10^{-4}$	$7.89\cdot10^{-2}$	$1.39\cdot10^{-5}$	$1.17\cdot10^{-3}$	$2.46\cdot10^{-5}$	$2.49\cdot10^{-4}$
IQR	$1.98\cdot10^{-4}$	$2.93\cdot10^{-1}$	$1.18\cdot10^{-5}$	$1.33\cdot10^{-3}$	$2.31\cdot10^{-5}$	$1.26\cdot10^{-3}$
Winsor. (Q^{05th})	$2.80\cdot10^{-4}$	$1.13\cdot10^{-2}$	$1.08\cdot10^{-5}$	$1.51\cdot10^{-3}$	$5.73\cdot10^{-6}$	$2.66\cdot10^{-3}$
Linear	$1.53\cdot10^{-2}$	$1.80\cdot10^{-4}$	$9.26\cdot10^{-2}$	$2.35\cdot10^{-2}$	$1.43\cdot10^{-2}$	$8.88\cdot10^{-3}$
Trim. (Q^{01th})	$8.83\cdot10^{-2}$	$5.51\cdot10^{-2}$	$2.31\cdot10^{-1}$	$9.85\cdot10^{-1}$	$2.65\cdot10^{-1}$	$6.89\cdot10^{-2}$
Trim. (Q^{05th})	$9.25\cdot10^{-2}$	$5.55\cdot10^{-2}$	$3.26\cdot10^{-1}$	$8.48\cdot10^{-1}$	$1.74\cdot10^{-1}$	$7.20\cdot10^{-2}$

All 11 normalization strategy significantly improved segmentation accuracy in respect to the relative overlap to the manual traces. Paired t-test is conducted for all 11 normalization method against the one without normalization, denoted as *None* from the results of 33 paired segmentation results for all six subcortical structures: nucleus accumben *(accu)*, caudate nucleus *(caud)*, globus pallidum *(glob)*, hippocampus *(hipp)*, putamen *(puta)*, and thalamus *(thal)*. 11 normalization transformation includes double sigmoid *(D.Sig.)* $\alpha=1$ and 5, sigmoid *(Sig.)* $\alpha=1$ and 5, z-Score, min/max, IQR based linear transform, Winsorizing *(Winsor.)* $\alpha=1$ and 5, and Trimming *(Trim.)* $\alpha=1$ and 5. Also note that the methods are organized from top to bottom based on six structures' average performance with regard to ICCs.

4.4 Discussion

Impact of Region-Specific Normalization: Experiments showed that the region-specific intensity normalization enables a successful segmentation of subcortical structures across a wide range of input image characteristics with as much accuracy

Table 4.4: Statistical significant of all 11 normalization variation against Double Sigmoid ($alpha = 1$), one of the best performed normalization transform

method	*accu*	*caud*	*glob*	*hipp*	*puta*	*thal*
D.Sig(Q^{05th})	$5.28{\cdot}10^{-1}$	$3.59{\cdot}10^{-1}$	$1.71{\cdot}10^{-2}$	$4.41{\cdot}10^{-1}$	$5.74{\cdot}10^{-1}$	$3.23{\cdot}10^{-1}$
zScore	$4.93{\cdot}10^{-1}$	$1.76{\cdot}10^{-1}$	$1.72{\cdot}10^{-1}$	$1.29{\cdot}10^{-2}$	$2.97{\cdot}10^{-2}$	$8.35{\cdot}10^{-3}$
Sig.(Q^{05th})	$9.64{\cdot}10^{-1}$	$1.01{\cdot}10^{-3}$	$3.21{\cdot}10^{-1}$	$8.55{\cdot}10^{-1}$	$7.47{\cdot}10^{-1}$	$8.67{\cdot}10^{-3}$
Winsor.(Q^{01th})	$1.13{\cdot}10^{-1}$	$5.67{\cdot}10^{-1}$	$6.96{\cdot}10^{-2}$	$4.56{\cdot}10^{-2}$	$2.22{\cdot}10^{-1}$	$1.78{\cdot}10^{-2}$
Sig.(Q^{01th})	$5.14{\cdot}10^{-1}$	$1.21{\cdot}10^{-3}$	$2.88{\cdot}10^{-1}$	$5.55{\cdot}10^{-1}$	$7.26{\cdot}10^{-2}$	$3.62{\cdot}10^{-2}$
IQR	$9.11{\cdot}10^{-1}$	$1.51{\cdot}10^{-3}$	$2.53{\cdot}10^{-1}$	$8.86{\cdot}10^{-1}$	$3.39{\cdot}10^{-1}$	$1.66{\cdot}10^{-4}$
Winsor.(Q^{05th})	$2.46{\cdot}10^{-1}$	$1.51{\cdot}10^{-1}$	$4.50{\cdot}10^{-1}$	$3.88{\cdot}10^{-2}$	$6.59{\cdot}10^{-1}$	$2.36{\cdot}10^{-3}$
Linear	$8.70{\cdot}10^{-3}$	$2.05{\cdot}10^{-1}$	$4.98{\cdot}10^{-8}$	$2.97{\cdot}10^{-4}$	$2.58{\cdot}10^{-5}$	$2.75{\cdot}10^{-7}$
Trim.(Q^{01th})	$3.02{\cdot}10^{-6}$	$5.38{\cdot}10^{-2}$	$1.32{\cdot}10^{-7}$	$5.04{\cdot}10^{-4}$	$4.33{\cdot}10^{-8}$	$1.82{\cdot}10^{-8}$
Trim.(Q^{05th})	$1.14{\cdot}10^{-6}$	$5.56{\cdot}10^{-2}$	$1.10{\cdot}10^{-7}$	$3.06{\cdot}10^{-4}$	$7.17{\cdot}10^{-8}$	$1.08{\cdot}10^{-8}$
None	$1.57{\cdot}10^{-4}$	$3.42{\cdot}10^{-3}$	$2.77{\cdot}10^{-5}$	$1.03{\cdot}10^{-3}$	$2.26{\cdot}10^{-5}$	$1.62{\cdot}10^{-4}$

Statistical significant of all 11 normalization variation against Double Sigmoid ($alpha = 1$), one of the best performed normalization transform. As methods are organized based on their performance, note that statistical differences from Double Sigmoid ($alpha = 1$) to both method at the top two rows, double sigmoid ($alpha = 5$) and zScore based linear transform, are the smallest among all other methods.

as manual methods. We believe that the success is primarily due to adequately enhanced structural boundaries in the scene while preserving consistency of biological contents across scans. For the in-vivo application, we have carefully selected two normalization $\mathcal{N}$ approaches after considering individual performance: *IQR*-based and *linear (min/max)*. The *IQR*-based normalization strategy was primarily used for its computational simplicity except for caudate nucleus; *linear (min/max)* transformation function was handpicked only for caudate nucleus due to its exceptional performance. Associated performance variations are reported in Figure 4.3.

Relative Spatial Location of Caudate: One should note that only caudate nucleus showed the best performance with the *linear (min/max)*. For the rest of ROIs the segmentation accuracy rather unacceptable with the *linear (min/max)* (Figure 4.3). That is, only caudate nucleus results in distinguished behavior: superior performance with the $linear(min/max)$ normalization. One explanation why caudate nucleus displayed different behavior than others could be found from its relative spatial location against neighbored tissues.

For the candidate nucleus candidate region, characteristics of the computed statistics are distinguished from other (Figure 4.4): 1) minimum statistics (red box) is more stable (a thin box) and 2) mean, median, and quantiles are more unreliable (a wider boxes) than other structures. It is obvious that, from the statistics as described above, caudate nucleus is better described by its minimum and maximum. The caudate nucleus is, therefore showed the best performance with the normalization function using min/max values, linear(min/max).

The different statistics of the caudate nucleus region is due to its relative spatial location. Caudate nucleus is adjacent to all three tissue types, including WM, GM, and CSF. Meanwhile, the rest of five subcortical ROIs shares boundaries only to the WM or other GM structures, but CSF. That is, the descriptive statistics of caudate nucleus region is under the CSF influence, which presents exceptional dark and bright intensity in I_{T1} and I_{T2} , respectively. In addition, the portion of CSF that is involved in the statistic computation also depends on the brain morphology and registration performance in the candidate-region identification step. As a result, the quantiles of the caudate nucleus region are to be unstable.

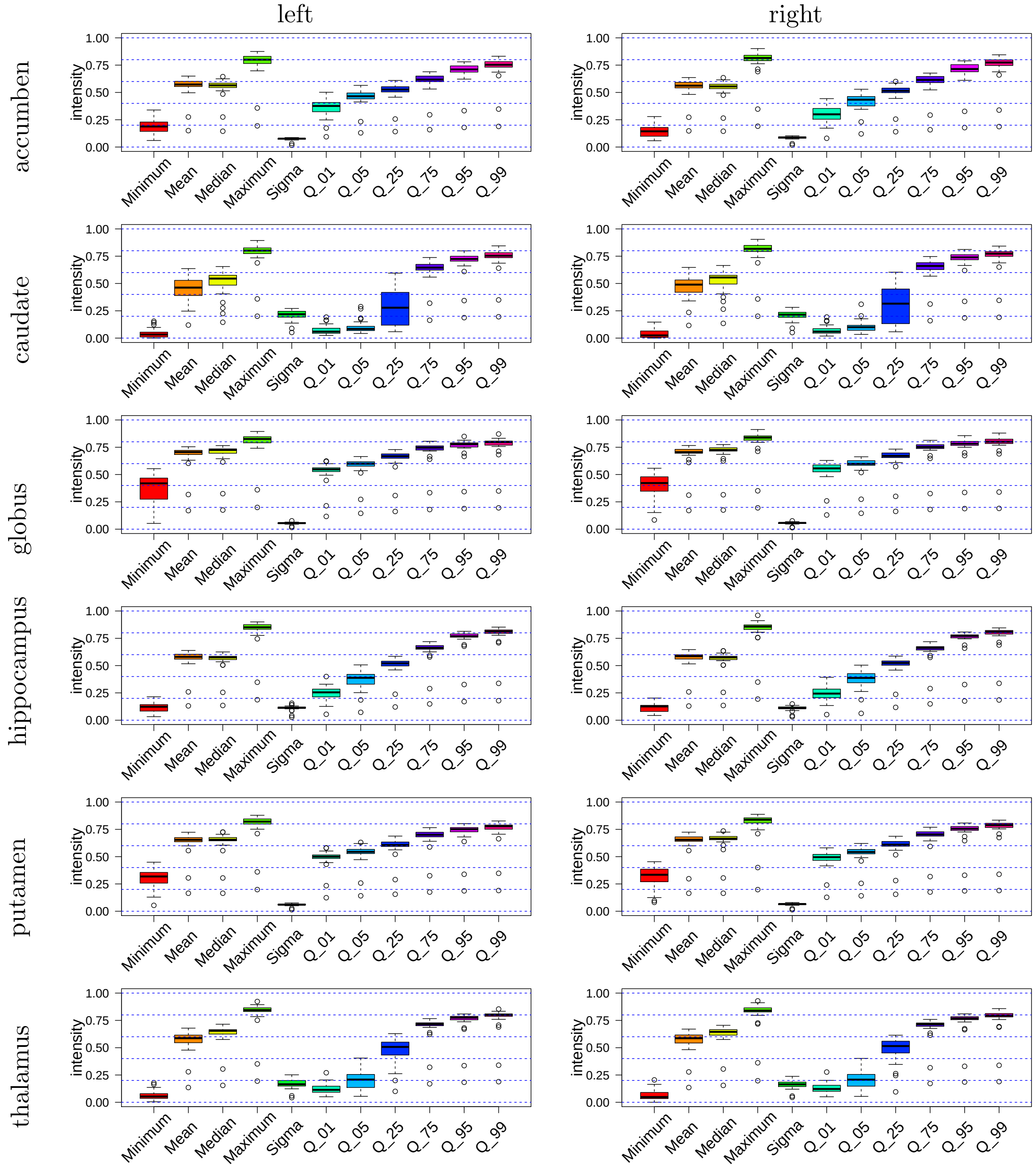


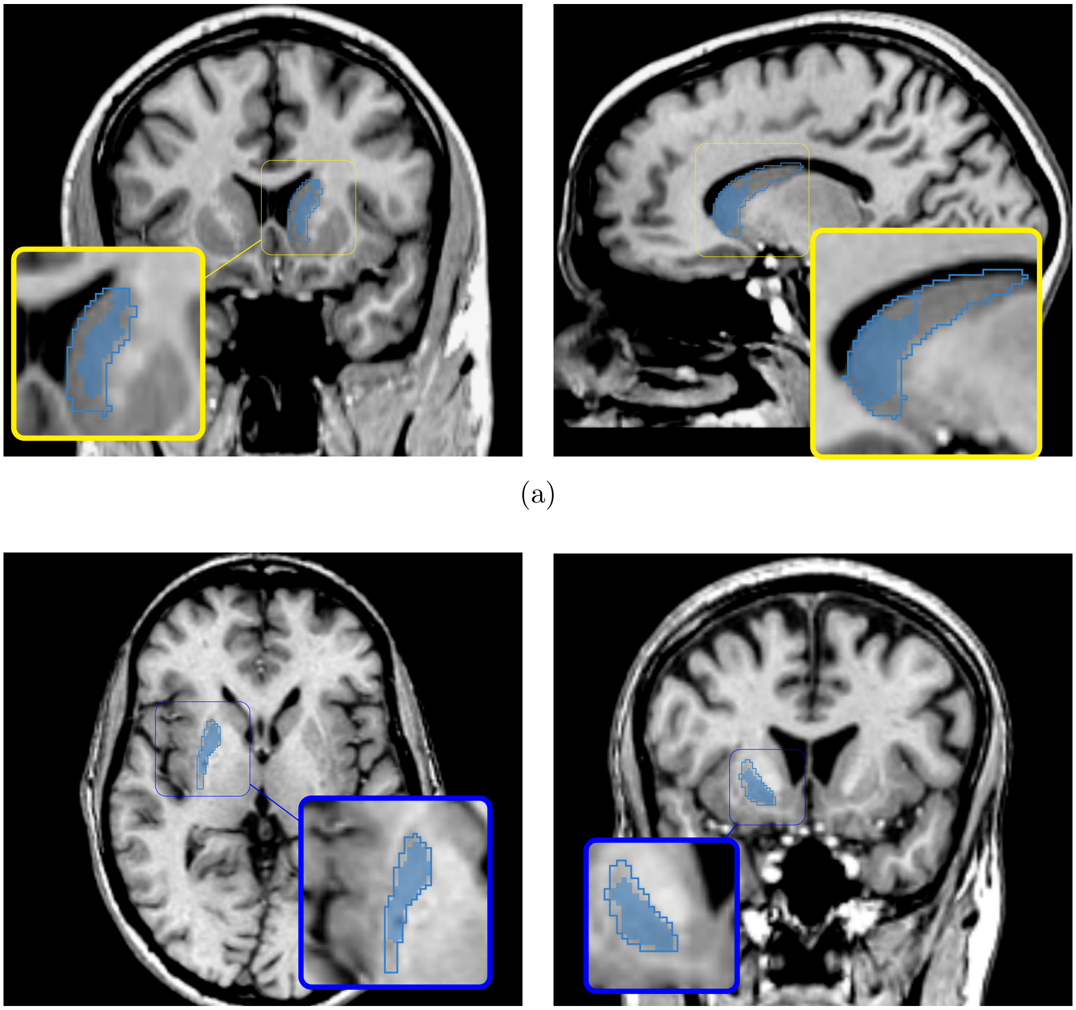
Figure 4.4: Multiple region specific statistics of intensity scaled to between $(0, 1)$ and computed in the candidate regions of interest identified by deformed priors. In general, each statistics shares similar trends but 25th quantile of caudate nucleus shows large variation across scans (wider blue box than other structures).

Visiting Failure/Success Cases Depending on the Choice of Normalization: As we mentioned above, each structures of interest have its own favorable normalization strategy (Figure 4.3). From our years of experience in the large-scale data processing, we know that there is always possibility for an algorithm to fail even with a very careful design and testing. To our knowledge, there is no shortcut to identify failures or outliers and improve the method other than exposing algorithm to real-world data. There is no *'One-Size-Fits-All'* solution. It is also important to try out the method on outside development data to ensure the method carries on applicable parameter set for real-world data.

To give some more intuitive sense of how each structure segmentation results greatly depends on the choice of normalization function, here we provide failure and success samples of caudate nucleus and putamen. As we discussed earlier, caudate nucleus was the best coupled with the linear (min/max) while other ROIs, including nucleus accumben, putamen, globus pallidum, thalamus, and hippocampus, do not work well with it. Relevant examples are provided in Figure 4.5. The Figure 4.5 shows that both success (outlined) and failure (filled) of caudate nucleus and putamen from two independent scans. By using linear (min/max) normalization approach, the segmentation algorithm succeeded for caudate nucleus (outlined in Figure 4.5a) but failed for putamen (filled in Figure 4.5b). Similarly, the segmentation approach with IQR-based normalization excellently segmented the putamen (outlined in Figure 4.5b) but underestimated the caudate nucleus (filled in Figure 4.5a). From these cases, it is obvious that different normalization method have a substantial effect on the segmentation results visually and quantitatively. This visually unpleasant segmentation results did not observed in our 32 training data set, even with 10-fold cross validation study. This scan has randomly been tried and identified as a failure by visual inspection process.

4.4.1 Summary & Conclusion

We evaluated potential benefits of total 11 intensity normalization strategies to our brain subcortical segmentation framework. The 11 variations of intensity normalization are originated from five main transform functions. For evaluation criteria, segmentation accuracy measured in $ICC(A)$ and $ICC(C)$ via 10-fold cross validation were employed. A series of contrasting experiments revealed significant benefits of all 11 region-specific normalization approaches. All the improvement achieved by 11 normalizations were statistically valid with some degree of variation (See Table 4.3). Of 11 normalizations, the best normalization strategy of *Double Sigmoid (alpha* $= 1$*)* resulted in the best performance on average comparing to rest methods. One should note that robust choice of normalization may differ based on disease state or scanning characteristics for a particular data set. In the our segmentation framework, it is obvious that the any choice of normalization strategies other than trimming advances the segmentation performance. The proper choice of function and statistics for region-specific normalization secured the excellent robustness level of our segmentation framework.



(a)

(b)

Figure 4.5: Failure and Success of Segmentation for left caudate nucleus 4.5(a) and right putamen 4.5(b). Two normalization methods are utlized on this example: 1) linear (min/max) and 2) IQR-based normalization. When linear (min/max) normalization is employed in the segmentation framework, the segmentation algorithm failed for the caudate nucleus but succeeded for the puatmen. On the other hand, the segmentation framework with the IQR-based normalization showed an excellent segmentation accuracy for the putamen, but underestimated the caudate nucleus. From these cases, we clearly see that different normalization method have a substantial effect on the segmentation results visually and quantitatively. This visually unpleasant segmentation results did not observed in our 32 training data set, even with 10-fold cross validation study. This scan has randomly been tried and identified as a failure by visual inspection process.

CHAPTER 5
FEATURE IMAGE SELECTION

A successful automatic segmentation tool requires the ability to identify and augment features that best characterize regions of interest consistently for a wide range of MR imaging characteristics. The quality of features is also crucial, especially for delicate structures of interest in human brain. We identified a set of features from the literature that includes primary edge-detectors, high-order image descriptors including statistical features [143, 88, 155], geometric moment invariant features [176], and multi-scale features [133, 118]. For this discovery feature set, we investigated a few subsets by effectively adapting a hierarchical feature forward selection approach. Through a series of comparative experiments, the gradient magnitude sum image from the T1- and T2-weighted images was found to be the best for the MR subcortical structures. We provide the detailed results of each experiment and discuss the effects of each feature-enhanced image on the subcortical segmentation accuracy.

5.1 Introduction

MR images hold a rich landscape of information about soft tissues in the human brain, but identifying the most relevant information is a complex endeavor. Although the relevant information is easily available to human experts, the process of delivering this wealth of information to machine-learning algorithms via feature-enhanced images is still a challenge. Informative feature-enhanced images effectively guide a ML-based tool to build a rigorous brain MRI segmentation model. This study is designed to identify the most advantageous feature-enhanced images to enhance a robust processing of multicenter large-scale MR data.

There exists several possible feature-enhanced images that describe the essence of the scene such as edges, sub-region or other salient characteristics, etc. These feature-enhanced images provide a higher level of semantics effectively to the algorithm, compared to raw, non-processed images. Of numerous feature images available in the field, one might naively assume that all possible images should be included to deliver as much information as possible. This approach, however, has drawbacks, for instance it increases computational complexity which can lead to intractable problems. Generally, feature selection in a supervised learning task aims to reduce the number of dimensions (features) considered in a task [1] while maintaining all the relevant information to improve performance and accuracy.

Correct feature selection is important for successful and efficient segmentation tool development. Furthermore, it was reported that the most machine-learning algorithms can be adversely affected by the input including irrelevant and/or redundant information [63]. Therefore, it is crucial to identify a adequate set of features that is minimally sized and maximally informative for a robust and accurate automated segmentation tool.

We discuss features extracted for brain subcortical segmentations from MR images and describe the process of feature image selection to improve performance of the segmentation framework.

5.2 Background

Machine-learning provides a powerful mechanism for making predictions, assuming the availability of training instances defining the learning task. In the machine-learning task, algorithms are provided a finite training set of labeled vectors, or instances from which to induce a predictive machine-learning model. This machine-learning model, in turn, is used to class label from unseen instances. Thus, in the building of a machine-learning model, the classification information that is inherent to the feature is of foremost importance [86]. We first visit feature subset selection method paradigm (Section 5.2.1) and followed by descriptions of feature-enhanced images used in this study (Section 5.2.2).

5.2.1 Feature Subset Selection

Feature subset selection attempts to reduce the number of dimensions considered in training to maximize the performance of the machine-learning algorithm. The goal of feature selection procedure is to identify the most relevant features while eliminating redundant and unrelated ones that can adversely affect on segmentation performance. Feature subset selection is an active research area, so we briefly introduce basic procedure that can be adapted to improve segmentation performance.

Feature selection should be employed so it effectively improves accuracy on its task. Theoretically and intuitively, it seems natural to utilize all the information at hands for the best prediction. In practice, however this approach suffers from the *curse of dimensionality*. That is, as the number of features in a induction (classification) task increases, computational complexity grows dramatically. It is often stated that when the set of features is sufficiently large, many machine-learning algorithms are simply intractable [86, 63]. In addition, the learning process is further exacerbated since many features may either be irrelevant or redundant to other features [86]. In this context, such features serve no purpose except to increase computational complexity [86]. Furthermore, feature selection is required when features are expensive to obtain or when one wants to extract meaningful relationship between features and class labels.

In general, a feature selection procedure requires two main components: 1) an objective function to accurately evaluate these candidates and 2) a strategy to accordingly select the candidate subsets. Rest of sections describes the objective function of our choice in this study (Section 5.2.1.1) and a approach to effectively drive a proper feature subset from the available feature domain (Section 5.2.1.2).

5.2.1.1 Objective Function

In general, an objective function could be drawn from two types: a filter type or a wrapper type. The filter type evaluates each feature subset according to its information content, e.g. interclass distance, statistical dependency, or information-theoretic measures. In contrast, the wrapper type directly utilizes a classifier model to evaluate a feature subset relying on prediction accuracy. Since the true prediction accuracy is usually unknown, the wrapper type estimates the prediction accuracy by means of statistical resampling such as boosting, or the hold-out method.

Advantages and disadvantages exist in the choice of objective function. Table 5.1 summarizes the characteristics of each type for feature subset selection. It is commonly observed that the filter type is more time efficient in comparison to the wrapper type more generalizable since it is built solely upon information presented in the training set, such as correlation and mutual information, independent of the machine-learning model. On the other hand, the result feature subset from the wrapper is optimally specific to the classifier under consideration [3] and is known to achieve better recognition rates, (i.e. prediction accuracy), because it is built upon the specific interactions between the classifier and the dataset. That being said, while the wrapper type could lead to an optimal subset for the classifier, the result subset may not be generalizable. The wrapper type, however, has a slower development process since the subset selection is based on the classifier model which requires the thorough complete machine-learning process containing all iterative procedures in training and testing.

5.2.1.2 Sequential Forward Subset Selection (FSS) of Features

One of commonly used approach in feature selection is the sequential forward subset selection (sequential FSS) strategy. The basic procedure of the sequential FSS method is intuitive and simple to understand. The sequential FSS procedure [131] is as following:

Table 5.1: Properties of Objective Functions: Filter vs. Wrapper Type

Type	Example functions and Pros (+) & Cons (−)
Filter	Correlation coefficient, distance metric, interclass distance, etc. (+) Faster cycle of testing (+) Generality of result feature set (−) Possible large subset selection
Wrapper	Machine-learning model itself (−) Slower testing cycle by building entire machine-learning model (−) Lack in generality of result feature set (Very specific to the target wrapper) (+) Generally resulting in the model with higher prediction accuracy

Algorithm 3: Feature forward selection

```
Input: feature set$S$
Output: selected feature set $O$
Start with the empty set $Y_0 = \emptyset$ ;
Select the next best feature $x^+ = argmax_{x \notin Y_k} J(Y_k + x)$.;
Update $Y_{k+1} = Y_k + x^+; k = k + 1$.;
Go to 2, checks the improvement. ;
```

Another popular feature selection approach is *backward sequential selection* (BSS), which can outperform sequential FSS in certain cases. In contrast to FSS, BSS begins with all features and repeatedly removes features when removal yields a performance improvement [1]. In our work we choose FSS based on the need to minimize both memory and processing resources in the development process. Due to the volume of our data, increasing the number of involved volume leads to cumulated memory usage and processing time, which in turn, adversely affect on a prompted development process due to time resources. The interested reader can refer to [1] for a detailed description and discussion between BSS and FSS.

5.2.2 Feature-Enhanced Images

Our feature set domain of feature-enhanced images, provides information to enable our algorithms to segment regions of interest based on the quality of the description. Information provided to the algorithm is usually limited to a local (pixel) level description of object rather than global level. Figure 5.1 shows an example of the discrepancy between local and global information perceived by machine and human observers respectively: from the left to the right, the Figure 5.1 shows MR brain image, zoomed-in views of a local scene, and a MR brain image with expert's traces. Edges of objects are not obvious to a human observer from the local view on the right hand side. This information gap between algorithm and expert leads to a failure of automatic delineation of desired objects. Feature images , including edges, brightness, texture, and shape distributions, can fill the *'semantic gap'* between what humans perceive as an object and what objects the algorithm can recognize.

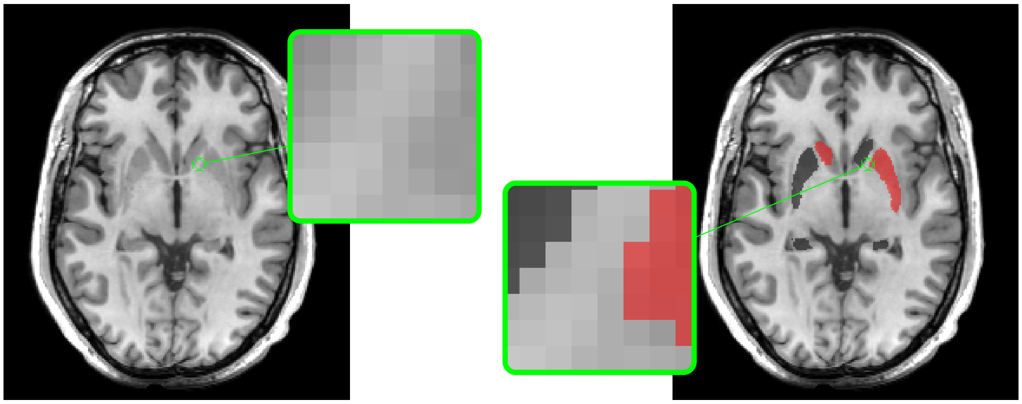


Figure 5.1: Semantic gap between information what human expects and what machine gets. The figure shows a MR brain (5.1a), MR brain with expert traces (5.1b), and zoomed-in views highlight the local information at the edge. Objects are identifiable to a human at the global view but it is too ambiguous to determine a precise boundaries from the localized scene. The algorithms are usually trained on the local information, where no obvious boundaries may exists.

To maximize the semantic information given to our segmentation algorithm, we have identified a number of features of brain MR images from literature and visually inspected them for its practical usefulness. Types of identified feature-enhanced images are sub-divided into **“edge-enhanced group”** and **“descriptive-subregion group”**. A subset of identified features are shown in Figure 5.3 and descriptions are given in the rest of this section.

To acquire robust feature-enhanced images, an anisotropic diffusion filter has been applied both on T1 and T2 modalities (denoted $\mathcal{I}_{T1}$ and $\mathcal{I}_{T2}$, respectively) and tested for their contribution to our segmentation. There are strong indications that the first stage of the human visual perception system has a large set of filter banks, where not only filters at multiple scales but also multiple orientations are present [15]. Perona et al.[117] presented another Gaussian smoothing operator, which compute coarse *'semantically meaningful'* edges of the scene: *anisotropic diffusion filter*. These denoising step with anisotropic diffusion filter are used as a *pre-processing* step to reinforce the processing robustness by minimizing noise in scans rather than to provide a new set of feature images.

5.2.2.1 Edge-enhanced Feature Images

Gradient magnitude ($\mathcal{I}_{GM}$) A gradient magnitude detects the edges at which pixels change their gray-level relatively suddenly. A gradient magnified images from $\mathcal{I}_{T1}$ MRI is shown in Figure 5.3a. Sum of gradient magnitude image of those $\mathcal{I}_{T1}$ and $\mathcal{I}_{T2}$ images computed according to the Equ. 5.1 to incorporate the gradients information from both multi-modal images.

$$GM(T1, T2) = |\bigtriangledown f_{T1}| + |\bigtriangledown f_{T2}| \tag{5.1}$$

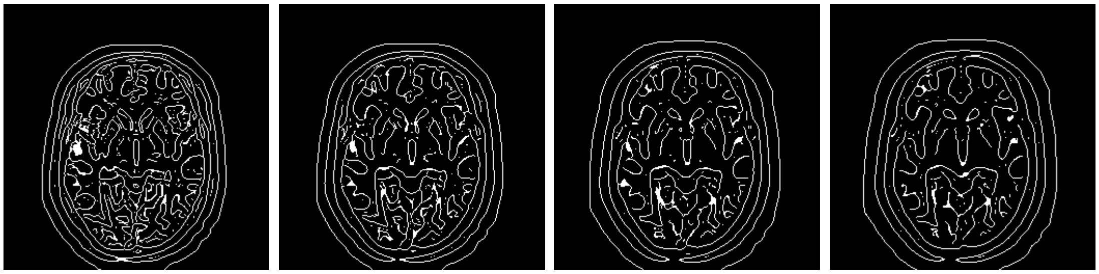
$$|\bigtriangledown f| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \left(\frac{\partial f}{\partial z}\right)^2} \tag{5.2}$$

Sobel operator [148] ($\mathcal{I}_{Sobel}$) One of the most traditional and popular edge detecting filter, Sobel operator, is also employed for this comparative test. Sobel operator also detects edges and example is given in Figure 5.3(b).

Multi-scale Edge ($\mathcal{I}_{E_m(I)}$) It is known that edges could be varied in different resolutions as Figure 5.2 shows edge-detected images at different scales. We could expect that the edge information at a coarser scale brings out global information while one at a fine scale underlines sufficient details of the scene. We combined multiple edge-detected images at different scale to effectively identify edges at global and local level for our region of interest. We have produced a combined edge image specifically by using Canny edge detection algorithm.

$$E_{multiScaleEdge} = \sum_i e_{\sigma_i},$$

where e_{σ_i} is a canny edge detected image at a standard deviation σ level of i. Figure 5.3(d) show an edge-combined image from Canny with $\sigma = 3, 6, 9$, and 12 respectively. For readers who are interested in detailed aspects of multi-scale features and their justification are well described in [133, 118].



(a) T1 $\sigma = 3$ (b) T1 $\sigma = 6$ (c) T2 $\sigma = 9$ (d) T2 $\sigma = 12$

Figure 5.2: Edges detected at different scale is known to deliver fine to coarse boundary information. The figure shows an example of $\mathcal{I}_{T1}$ filtered with canny edge filter of $\sigma = 3, 6, 9$ and $\sigma = 12$. In this example, the finest edges are detected at scale $\sigma = 3$ (Figure (a)) and the crudest edges are shown with $\sigma = 12$ (Figure (d)). Feature images that collaborated edges detected at multiple scales are provided as a candidate feature-enhanced image for our segmentation framework.

5.2.2.2 Neighborhood Statistic-based feature images

Mean ($\mathcal{I}_{Mean(I,r)}$) Mean filtered images with radius $r = 3$ was selected for testing. Figure 5.3(e) show mean filtered images with $r = 3$ applied at $\mathcal{I}_{T1}$ and $\mathcal{I}_{T2}$ MRI.

Median ($\mathcal{I}_{Median(I,r)}$) Median of neighborhood size with $r = 3$ was computed. Median is a locational parameter in descriptive statistics indicating a very middle of the value between neighbors. Figure 5.3(f) shows examples of median filtered image with $\mathcal{I}_{T1}$ and $\mathcal{I}_{T2}$.

Standard Deviation ($\mathcal{I}_{Std(I),r}$) Standard deviation of sub-region images are one of simple but not very commonly used in the image processing since it is not intuitive at first sight. The computation is very similar to mean or median images above but just calculates for standard deviation instead of mean or median in the sub-region. Standard deviation could be thought as an inhomogeneity measure of a selected region. It often called as a texture descriptor as it present a property of region. Standard deviation feature image computed as a candidate, where neighbors are limited by $r = 3$ (Figure 5.3 (g)).

5.2.2.3 Higher Level Semantic feature images

Geometric moment invariants ($\mathcal{I}_{GMI}$) Geometric moment invariants have successfully used in some medical image processing, especially registration. We have computed GMIs as candidates (See Figure5.3 (c)).

Posterior from BRAINSABC ($\mathcal{I}_{Posterior(Tissue)}$) BRAINSABC, which is a bias field correction module in BRAINS Tool, computes posterior probabilities of brain tissues: grey matter, white matter, and cerebrospinal fluid (CSF). The example of sub-cortical grey matter posterior is shown in the Figure 5.3(h).

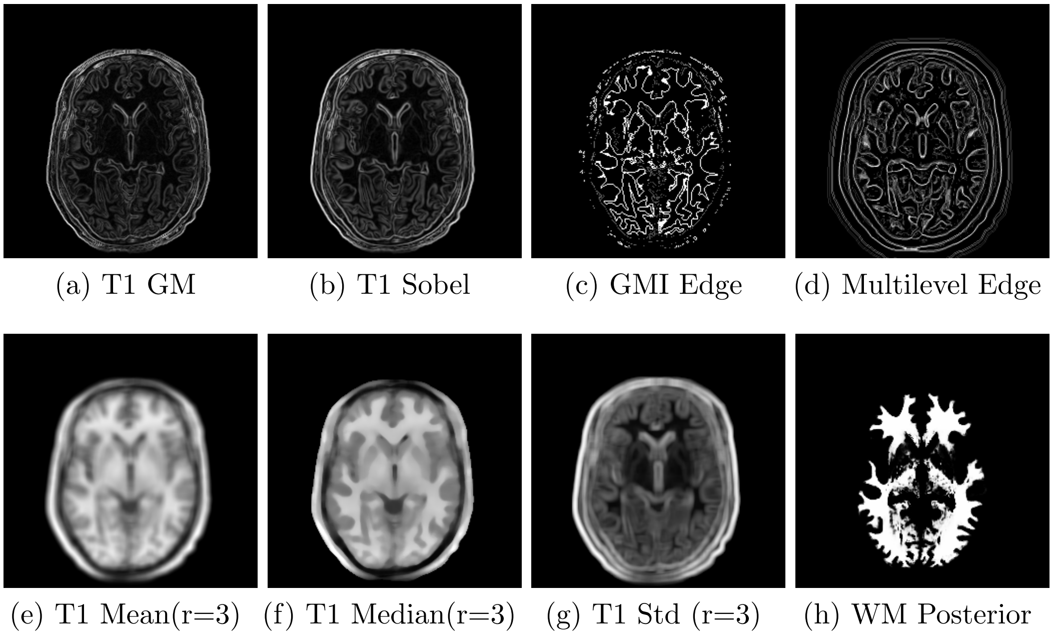


Figure 5.3: Feature-enhance images example used in this study for their effect on segmentation accuracy for subcortical structures. Edge-enhanced images are placed at the upper row in the Figure (5.3(e) 5.3(h)) and region-enhanced images are shown at the bottom row (5.3(e) 5.3(h)). Feature-enhanced images are commonly incorporated as input for the automated segmentation framework to provide semantically meaningful information about voxel.

5.3 Experiment and Result

To build a custom subset of feature-enhanced images efficiently and effectively, we have employed a hierarchical 'sequential feature forward selection' (FSS) approach. First, we divide all the identified feature-enhanced images into two sub-groups based on context of feature information: **edge-enhanced group** and **descriptive-subregion group**. For instance, while the first group includes all the primitive edge-enhanced images, the second group, *descriptive-subregion image* group, encompasses the statistic-based feature images that are computed on a neighborhood area and therefore carries higher level semantic. Upon both experimental subgroup of feature images, we conducted a 10-fold cross validation to compare segmentation accuracy across feature-enhanced images for large-scale multicenter data. Following Table 5.2 shows the corresponding feature subset that are used in this paper.

The comparative experiments suggests two feature-enhanced images that displayed an superior performance in terms of ICCs: gradient magnitude ($\mathcal{I}_{GM}$) and standard deviation ($\mathcal{I}_{Std,3}$) as shown in Figure 5.4. Those two candidates are one from each group of the edge-enhanced group and the descriptive-subregion group, respectively. The reported ICCs are averaged across four intensity normalization approaches, including double sigmoid ($\alpha = 01$), IQR, Linear, as well as no normalization and plotted adequately in decreasing manner from top to bottom. For the both group, while all eight features improve the segmentation accuracy to some extent, the degree of gain decreases as following:

edge-enhanced group: $\mathcal{I}_{GM} > \mathcal{I}_{Sobel} > \mathcal{I}_{E_m(T1)} > \mathcal{I}_{GMI}$

descriptive-subregion group: $\mathcal{I}_{Std(T1,3)} > \mathcal{I}_{Mean} > \mathcal{I}_{Median} > \mathcal{I}_{Posterior(WM)}$.

Further details of independent performance comparison of the four intensity normalization experiments are also given in Appendix A.4.

Table 5.2: Feature Subset List

Acronym	a set of images that are used
original image of T1 and T2:	
Raw_T1	$\{\mathcal{I}_{RawT1} \vert \text{AC-PC Aligned T1}\}$
Only	$\{\mathcal{I}_{T1} \vert \text{Bias correctied T1}\}$
T2	$\{\mathcal{I}_{T1}, \mathcal{I}_{T2} \vert \text{Bias corrected T2}\}$
subset of edge features:	
SG	$\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1,T2)}\}$
GM	$\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1)}\}$
Sobel	$\{\mathcal{I}_{T1}, \mathcal{I}_{Sobel(T1,3)}\}$
GMEEdgeInformation	$\{\mathcal{I}_{T1}, \mathcal{I}_{GMI}\}$
subset of region information features:	
MultiCannyT2S5	$\{\mathcal{I}_{T1}, \mathcal{I}_{E_m(T1)}\}$
Mean	$\{\mathcal{I}_{T1}, \mathcal{I}_{Mean(T1,3)}\}$
Std	$\{\mathcal{I}_{T1}, \mathcal{I}_{Std(T1,3)}\}$
Median	$\{\mathcal{I}_{T1}, \mathcal{I}_{Median(T1,3)}\}$
WM_Posterion	$\{\mathcal{I}_{T1}, \mathcal{I}_{Posterior(WM)}\}$
subset of two feature images:	
SG_Std	$\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}\}$
T2_SG	$\{\mathcal{I}_{T1}, \mathcal{I}_{T2}, \mathcal{I}_{GM(T1,T2)}\}$
T2_Std	$\{\mathcal{I}_{T1}, \mathcal{I}_{T2}, \mathcal{I}_{Std(T1,3)}\}$
subset of three feature images:	
T2_SG_Std	$\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}\}$

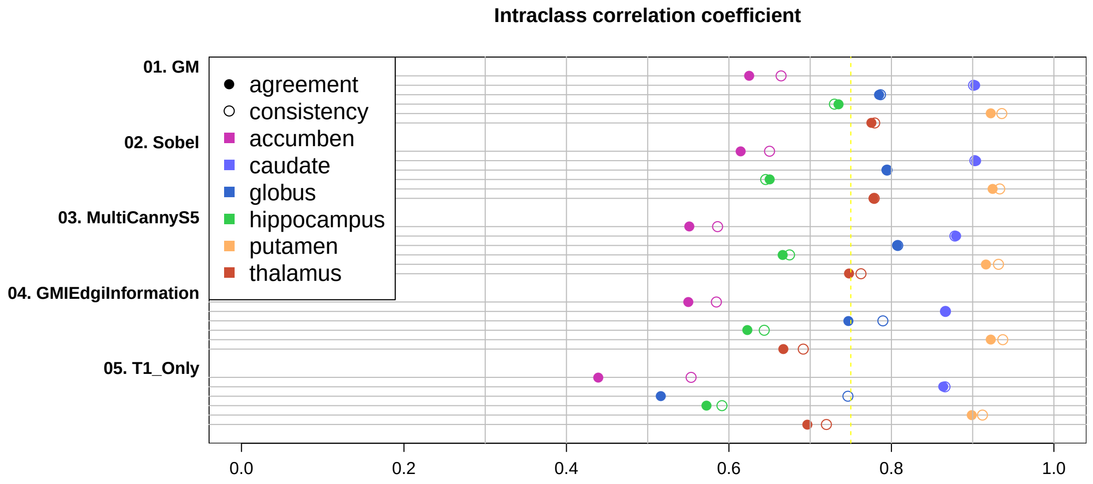


Figure 5.4: Comparative study for Feature Subset Image Selection: **edge-enhanced group**: ICC for edge-enhanced images. For four edge-enhanced images of , $f \in \{\mathcal{I}_{Sobel}, \mathcal{I}_{E_m(T1)}, \mathcal{I}_{GMI}, \mathcal{I}_{GM(T1)}\}$ contribution of each images to the segmentation framework are compared from the subset $\{\mathcal{I}_{T1}, \mathcal{I}_f\}$ in terms of average ICC across four intensity normalization approaches. Results are ordered according to its performance rank from top to bottom. The figure highlights the most beneficial features of $\mathcal{I}_{GM(T1)}$ (**01.GM**) among those four feature images. The performance of segmentation based on $\mathcal{I}_{T1}$ only (**05.T1Only**) is also provide a baseline, which ranked at the last, showing the worst performance with no aid from feature images.

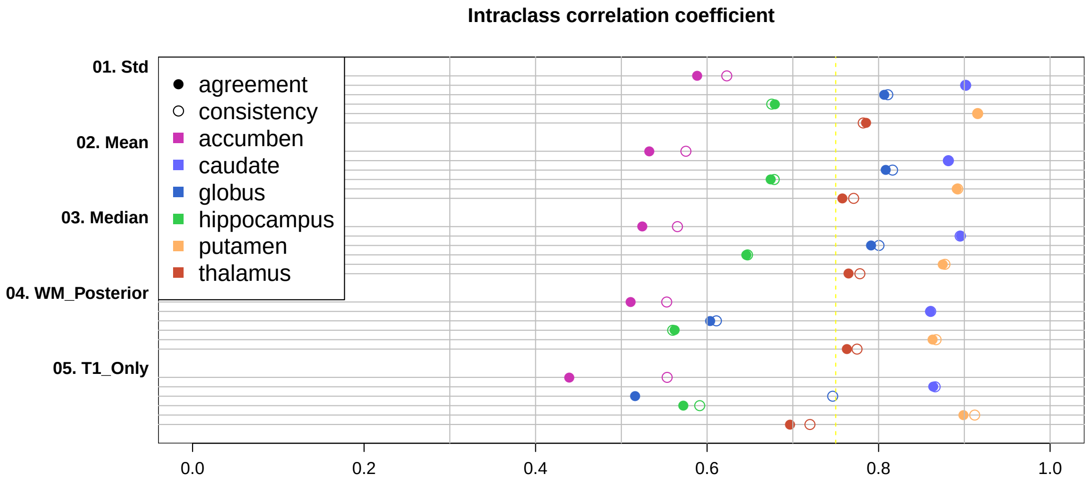


Figure 5.5: Comparative study for Feature Subset Image Selection: **descriptive-subregion group**. Paired to the Figure 5.4, contribution of feature images $\{\mathcal{I}_{T1}, \mathcal{I}_f\}$ from **descriptive-subregion group** (where $f \in \mathcal{I}_{Std(T1,3)}, \mathcal{I}_{Mean}, \mathcal{I}_{Median}, \mathcal{I}_{Std(T1,3)}, \mathcal{I}_{Posterior(WM)}\}$) to the segmentation framework is investigated and plots also ordered according to its performance rank. The graph highlight the most beneficial features of $\mathcal{I}_{Std(T1,3)}$ (**01.Std03**) in the descriptive-subregion group. The performance of segmentation based on $\{\mathcal{I}_{T1}\}$ only (**05.T1Only**) is also plotted to provide a baseline, which ranked at the last, showing the worst performance with no aid from feature images.

We continue the next level comparative experiments against the candidate feature images to finalize our custom feature subset. Two candidate feature-enhanced images, one from each of two groups, are identified by above experiments: 1) $\mathcal{I}_{GM(T1)}$ (denoted as GM) showed the best performance enhancement from the edge-enhanced group (Figure 5.4 and 2) standard deviation-based feature image $\mathcal{I}_{Std(T1,3)}$ (denoted as std) were the most promising candidate from the descriptive-subregion group, (Figure 5.4). The experiments are designed to evaluate performance of subset among four images; four images include two candidate features ($\mathcal{I}_{GM}$ and $\mathcal{I}_{Std(I,r)}$), and $\mathcal{I}_{T1}$ and $\mathcal{I}_{T2}$ images.

Experiment results are summarized according to the order of correspondence to manual traces of six subcortical structures as shown in Figure 5.6. As usual, ICC is served as correspondence measure between the automated segmentation and the gold standard (manual traces). In general, the subset that comprises $\mathcal{I}_{GM(T1,T2)}$ ranked at the top while the subset with both $\mathcal{I}_{GM(T1,T2)}$ and $\mathcal{I}_{Std(T1,3)}$ ranked at the second top. Any subset of features advances the segmentation accuracy on average, and the order is as following: $\{\mathcal{I}_{GM(T1,T2)}\} > \{\mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}\} > \{\mathcal{I}_{Std(T1,3)}\} > \{\mathcal{I}_{T2}, \mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}\} > \{\mathcal{I}_{T2}, \mathcal{I}_{GM(T1,T2)}\} > \{\mathcal{I}_{T2}, \mathcal{I}_{Std(T1,3)}\} > \{\mathcal{I}_{T1}\}$. One should also note that the inclusive subset of all four images are only ranked at 4 in this experiment and also $\mathcal{I}_{T2}$ adversely affects the segmentation performance.

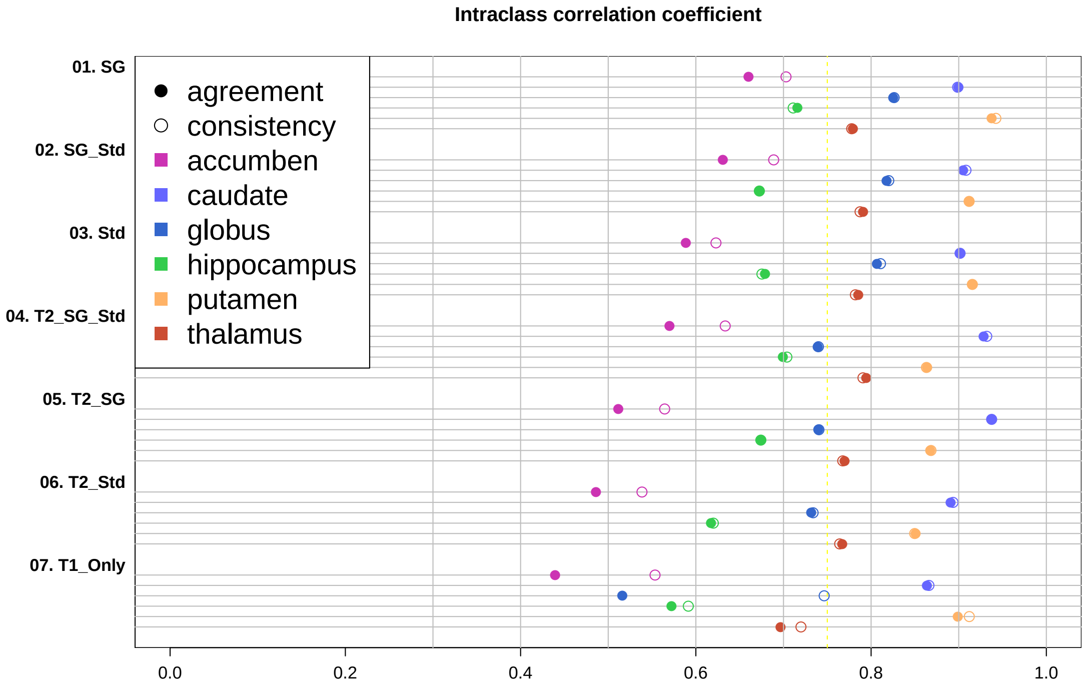


Figure 5.6: Segmentation accuracy contrasted for seven subsets $\{\mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}, \mathcal{I}_{T2}, \mathcal{I}_{T1}\}$ The results favors the subset $\{\mathcal{I}_{GM(T1,T2)}\}$ to the combined subset $\{\mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}\}$ in terms of average ICC over four different intensity normalization strategies: none, linear(min/max), IQR-based linear, double sigmoid ($\alpha = 1$) normalization (See Section 4 for normalization methods). Subset of features used here is selected based on the previous results performed $\{\mathcal{I}_{T1}\ ,\ \mathcal{I}_{feature_image}\}$, that identified the best performing features from each of **edge-enhanced group** and **descriptive-subregion group**. Note that experiment name is based on the features used in addition to $\mathcal{I}_{T1}$ and the method is ranked from the top to bottom as numbered in front of each experiment name. One should also note that the inclusive subset of all four images are only ranked at 4 in this experiment and also $\mathcal{I}_{T2}$ adversely affect the segmentation performance.

5.4 Discussion

Contribution of T2-weighted MRI to Segmentation Accuracy: Perhaps surprisingly, the discriminating power of T2-weighted MR image ($\mathcal{I}_{T2}$) were quite disappointing given that it is such a popular modality to collect together with $\mathcal{I}_{T1}$. Figure 5.7 clearly depicts how $\mathcal{I}_{T2}$ has an impact on segmentation results. At first glance, it is clear that feature set of $\{\mathcal{I}_{T1}, \mathcal{I}_{T2}\}$ are better than $\{\mathcal{I}_{T1}\}$, experiments denoted by *'13 T2'* and *'15 T1 Only'*, respectively. The rest of results, however, suggest that $\mathcal{I}_{T2}$ by itself may have an detrimental circumstances in terms of subcortical segmentation task:

- **01.** $\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1,T2)}\}$ vs. **11.** $\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{\mathbf{T2}}\}$,
- **04.** $\{\mathcal{I}_{T1}, \mathcal{I}_{Std(T1,3)}\}$ vs. **12.** $\{\mathcal{I}_{T1}, \mathcal{I}_{Std(T1,3)}, \mathcal{I}_{\mathbf{T2}}\}$,
- **03.** $\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}\}$ vs. **06.** $\{\mathcal{I}_{T1}, \mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std(T1,3)}, \mathcal{I}_{\mathbf{T2}}\}$.

All of above paired experiments point out the consequence of having $\mathcal{I}_{T2}$ into the feature subset. It is worthwhile to note that even though the most of results $\mathcal{I}_{T2}$the experiment with $\mathcal{I}_{GM(T1,T2)}$, which derived from both $\mathcal{I}_{T1}$ and $\mathcal{I}_{T2}$, ranked at the top. In other words, the information provided $\mathcal{I}_{T2}$ as it is may not be as beneficial as the one extracted through the gradient magnitude filter. The possible cause of this degraded segmentation accuracy when it involves $\mathcal{I}_{T2}$ is imperfect intra-subject registration between multi-modal scans.

Feature-enhanced Images VS. Region-Specific Intensity Normalization: To improve accuracy of segmentation framework, through out previous chapter (Chapter 5) and this one (Chapter 4), two main factors were investigated indepen-

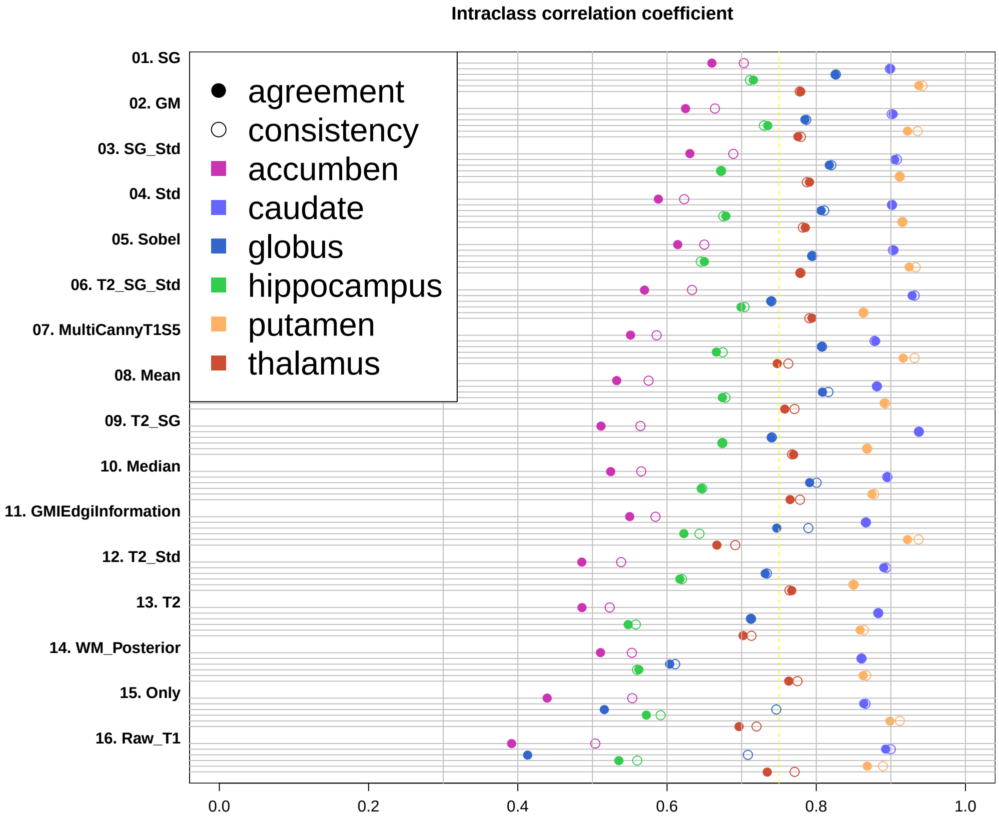


Figure 5.7: All feature-enhanced images are compared in one plot with ICCs averaged over four region-specific normalization strategies: none, linear(min/max), IQR, double sigmoid ($\alpha = 1$). The graph allows us to observe global tendency of each feature contribution to the segmentation accuracy in comparison to each other. As we concluded already, $\mathcal{I}_{GM(T1,T2)}$ advances the segmentation accuracy the most while any features were beneficial when it used together with $\mathcal{I}_{T1}$. Also note that performance contrasted between 15-17 shows how our pre-processing steps improves segmentation accuracy in terms of ICCs.

dently for their effectiveness: 1) feature-enhanced images (Chapter 5) and region-specific intensity normalization (Chapter 4). Our hypothesis was that those two factors would have collaborative effect on the segmentation framework. We have experimentally proved that those two factors are all practically useful to some extent. Here we discuss those two elements together, intensity normalization approaches and the choice of feature-enhanced images, with regard to their contribution to the segmentation framework.

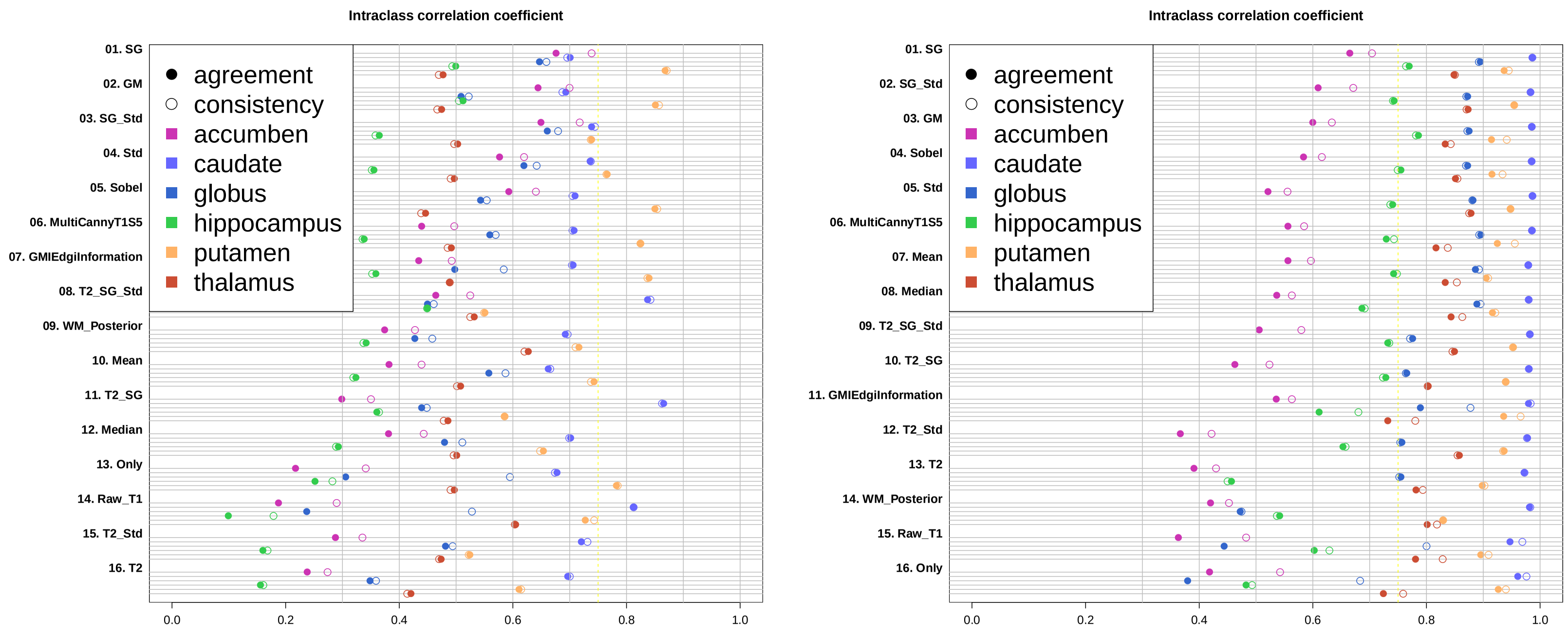
As discussed in the Chapter 4, in general, any region-specific normalization strategies are advantageous in terms of segmentation accuracy. In addition, any feature-enhanced images, additional to the $\mathcal{I}_{T1}$, also advanced segmentation accuracy in our framework, as well. The degrees of improvement, however, were varied to some extent. Figure 5.8 and Figure 5.9 shows segmentation accuracy contrasted across four selected normalization functions and subsets of features that we have tried:

Four Normalization Function Tested:

1. No normalization function,
2. Linear (min/max) normalization function,
3. IQR-based normalization function, and
4. Double sigmoid ($\alpha = 1$) normalization function

Feature-enhanced Images Tested:

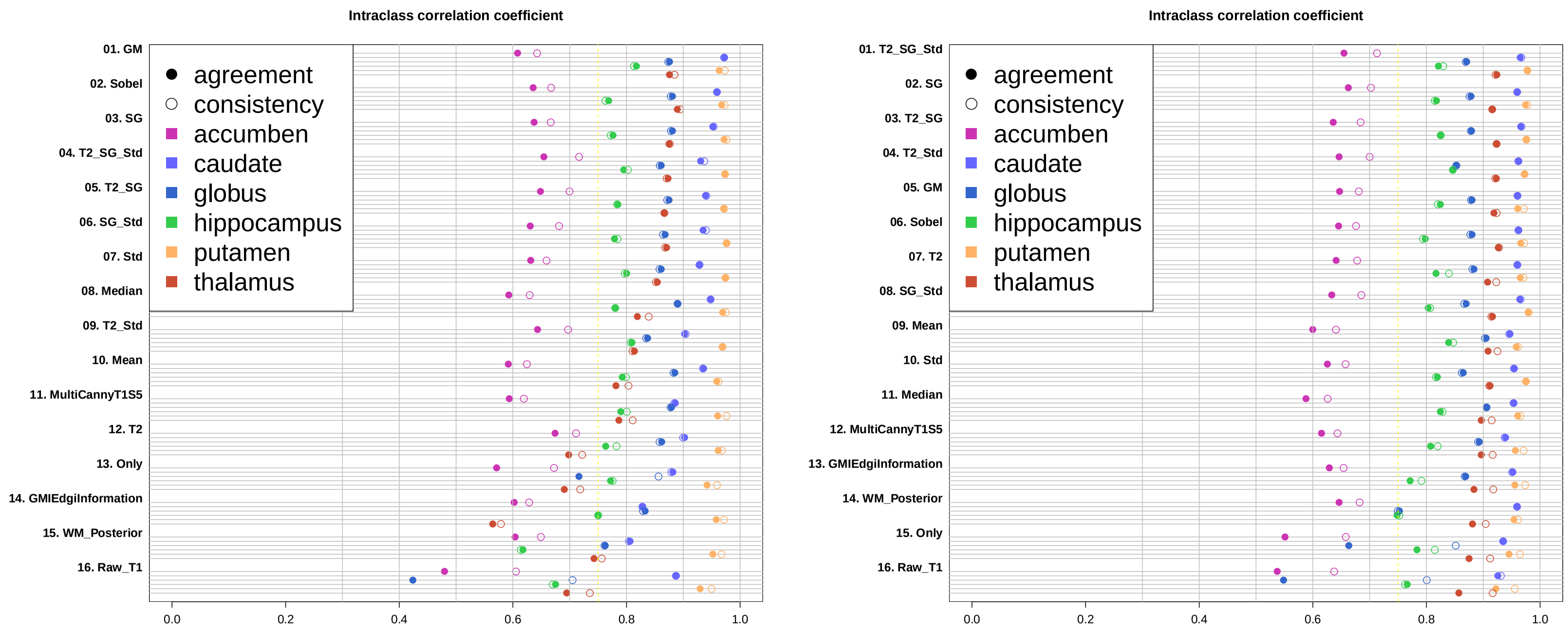
$$\{\mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std}, \mathcal{I}_{Sobel}, \mathcal{I}_{Mean}, \mathcal{I}_{Median}, \mathcal{I}_{E_m(T1)}, \mathcal{I}_{GMI}, \mathcal{I}_{T1}, \mathcal{I}_{RawT1}\}$$



(a)

(b)

Figure 5.8: Segmentation performance were contrasted with 16 different feature-enhanced images for four different normalization methodologies: no normalization (*None*) and linear (min/max) transform (*Linear*) is shown here and the other two is shown in Figure 5.9. The performance by using region-specific normalization with linear (min/max) transformation was shown noticeably consistent and better segmentation accuracy across any combination of feature-enhanced images as well as subcortical structures. Also note that there are subtle differences in ordering of feature set, where the ordering reflexes the contribution of feature set to the performance.



(a)

(b)

Figure 5.9: Segmentation performance were contrasted with 16 different feature-enhanced images for four different normalization methodologies: *IQR* and *double sigmoid* $(\alpha = 1)$ is shown here and the other two is shown in Figure 5.8. While two groups of experiments also present different feature preferences as appeared in the different ordering, 5.9(b) has more consistent performance trend across feature set tried than (a). The normalization strategy using *double sigmoid* $(\alpha = 1)$ is one of our top ranked normalization function in Section 4.2.2, which displayed the most consistent results across feature set experiments. The performance consistency across feature set is more noticeable when it is compared to Figure 5.8a.

To improve our understanding of sensitivity of two factors, the intensity normalization and the feature subset selection, the cumulative plots across four normalization strategies encompassing 17 feature subsets are presented in two places: Figure 5.8 and Figure 5.9. The results suggest feature sensitivity increases when no normalization strategy applied. That is, the region-specific intensity normalization desensitize the segmentation algorithm performance against the feature selection, therefore robust against choice of the features. The performance fluctuation is obviously severe without a normalization function (Figure 5.8a) comparing to other three sets of experiments of linear (min/max) (Figure 5.8b), IQR-based (Figure 5.9a), and double sigmoid (Figure 5.9b). We interpret this as following; once proper region-specific normalization is applied, the segmentation accuracy is less influenced by the choice of feature set that are used for the segmentation framework.

The primary design goal of region-specific normalization is to provide consistent intensity across scans while preserving biologically relevant information, i.e. boundaries, in the scene. As we discussed, automated tool development for robust and consistent subcortical segmentation are very challenging because of inter- and intra-scan intensity variation. In contrast to the machine, despite of those data inconsistency, trained experts often achieve accurate delineation with different scan types. We mentioned that the feature-enhanced images are commonly devised to compensate semantic gap between human expert and computer algorithm (See Section 5.2.2). Our experiments, however, suggested that intensity normalization may play an more important role in narrowing those semantic gaps for the machine-learning algorithm.

Some rationales about the importance of normalization step are also stated in *Weber's Law*:

> *The contrast sensitivity is approximately independent of the background luminance.* (Weber's Law)

Weber's Law implies that relative changes in luminance are important [19] and we can find partial explanation of where semantic gap is originated between what human experts perceive and what machine-learning algorithm is given. Human experts perceives regions of interest from the relative intensity changes in addition to their pre-attained knowledge while machine-learning struggles to learn about each structures based on absolute intensity values given as input (Figure 5.1). We believe that our region-specific normalization allowed our ML algorithm to narrow down the semantic gap by appreciating those relative changes. This was achieved through the intensity-based normalization with a proper choice of robust statistics.

5.4.1 Summary & Conclusion

To build a custom set of feature-enhanced images for large-scale multi-site data processing, wide variety of feature-enhanced images $\mathcal{F}$ are contrasted for their contribution to our proposed segmentation framework. For eight feature-enhanced images we identified, a series of experiments confirmed that the subset of $\{\mathcal{I}_{T1}, \mathcal{I}_{T2}, \mathcal{I}_{GM(T1,T2)}\}$ (or $\{\mathcal{I}_{T1}, \mathcal{I}_{GM}\}$ for uni-modal scan session) effectively advanced the segmentation accuracy the most. Instead of unrealistic exhaustive search of all the possible instances, this study followed a hierarchical FSS to find a custom subset of feature-enhanced images. The comparative experiments was sufficiently designed to narrow down the most profitable one among feature-enhanced images that carry similar information. This selection procedure is also allow to avoid a chance to include redundant feature images. We first sub-divided feature images into two sub-groups of features based on their information characteristics: the **edge-enhanced group** and the **descriptive-subregion group**. Then the performance of feature subset $\{\mathcal{I}_{T1}, \mathcal{I}_{FeatureImage}\}$ were contrasted within each group and the best feature-enhanced image from each group are chosen: gradient magnitude image ($\mathcal{I}_{GM}$ or $\mathcal{I}_{GM(T1,T2)}$ for multi-modal dataset) from the first group and standard deviation image (Std) from the second group. Those two feature-enhanced images, $\mathcal{I}_{GM}$ (or $\mathcal{I}_{GM(T1,T2)}$) and $\mathcal{I}_{Std}$, are then investigated further. We have tested multiple combination subgroups from $\{\mathcal{I}_{T1}, \mathcal{I}_{T2}, \mathcal{I}_{GM(T1,T2)}, \mathcal{I}_{Std}\}$ by investigation their contribution to the segmentation accuracy. Surprisingly enough, adding region-enhanced feature Std to the edge-enhanced feature image $\mathcal{I}_{GM(T1,T2)}$ did not advance segmentation results (See Figure 5.6). The best and the only feature-enhanced image for our segmentation framework turned out to be $\mathcal{I}_{GM(T1,T2)}$ (or $\mathcal{I}_{GM(T1)}$ for single modality data), which we have long been used internally.

CHAPTER 6
EVALUATION OF AUTOMATED SEGMENTATION METHOD FOR LARGE-SCALE MULTI-CENTER LONGITUDINAL MR DATA

This chapter evaluated our proposed segmentation framework, *BRAINSCut*, in terms of validity and reliability. Reliability is evaluated in three aspects: algorithmic reliability against various MR qualities, multicenter reliability by using traveling human phantom data (THP), and repeated measure reliability through scan-rescan data. We did not find any statistically significant differences between centers from the THP data when we measured six subcortical volumes by **BRAINSCut**. We also evaluated the sensitivity of measurement against degenerative disease status, represented by CAP group from PREDICT-HD study [92] as well as healthy controls. Out of four groups for each six subcortical structures (twelve individual structures in and right hemisphere), there was no statistically significant disease group effect on the measurement stability when controlling for age and gender, except for left caudate nucleus in high CAP group. *BRAINSCut* has been rigorously tested with 10-fold cross validation, and we also report a result of large-scale multicenter MRI data from PREDICT-HD study; the success rate, visual inspection results. Sample examples demonstrate the validity of the subcortical segmentation results. Each result is reported and discussed in-depth in this report.

6.1 Introduction

Neuroanatomical investigation requires validity and reliability of an automated segmentation method that targets understanding the functional-anatomical patterns of brain. Slight differences in imaging methods could have a considerable impact on the validity and reliability of morphometric measures. As noted in [76], the morphometric measurement from brain MRI can be affected by multiple factors, including hydration status of the subject, instrument related factors such as scanner manufacturer, field strength, head RF coil, magnetic gradient, pulse sequence and image processing method. In addition, quality may differ across brain structures due to variability in tissue intensity profiles and in modeling algorithms [175].

We engineered our framework explicitly for application to large-scale data. The process of developing and validating an automated segmentation algorithm both rely on reducing error. Empowered by high-throughput computing paradigm and workflow environment (NiPype), the present study demonstrates a series of evaluations of segmentation qualities that are coupled with the development process. We demonstrate how the entire pipeline, encompassing processing framework and validation, is constructed by using NiPype and high-throughput computing resources.

Two main aspects of segmentation quality, validity and reliability, are illustrated and employed to evaluate our proposed framework. Validity of segmentation quality is assessed with correspondence to the manual traces as well as with the visual inspections. Reliability of the segmentation is then characterized by employing two set of repeated scans: 1) traveling human phantom data and 2) scan-rescan data set.

The remainder of this paper is structured as follows: Section 6.2 reviews the recent validation study on large-scale MR data processing. Section 6.3 describes our software environment and resources that maximized our iterative and effective development procedure. Through Section 6.4 to Section 6.6 introduces evaluation data and model with in-depth discussion.

6.2 Related Work

In the literature, few studies have been conducted for the validation of MR brain processing collected at multicenter. The review study of this section is limited only to those brain MR processing assessment study for multicenter research.

One of earlier studies by Jovicich et al [77] reported for the reliability of their image distortion correction method on 1.5 Tesla multicenter MR data. The study suggested that image intensity reproducibility of the human MRI can significantly improved with their method and thus the method may offer improved reproducibility in morphometry studies. Fu et al. [49] conducted a multicenter MR comparative study and analyzed signal-to-noise ratio on those collected at multicenter but from a single vendor (GE). Multi-center MR data comparison was also conducted by Fu et al. [49] an restricted study for MR's signal-to-noise ratio analysis for the data collected from a single-vendor (GE). Their study stated that trends observed over time often depend on center and on modality and scanner manufacturer.

Gouttard et. al, [58] presented a traveling human phantom study for quantitative analysis of MRI including inter- and intra-center comparison. The study reports cross-center reliability, focusing on reproducibility, of automatic atlas-based segmentation results for five subcortical structures, including amygdala, caudate nucleus, hippocampus, pallidus (globus pallidus), and putamen in both left and right hemisphere. The study showed the detailed degree of segmentation reliability between MR data acquired from two centers with a unique MR vendor (Siemens).

A more complete reliability study of longitudinal and multicenter MRI data was also reported in [78]. Scan-rescan reliability is reported in [102] for a single site MR data with sample size estimation as well. Morey et al. [102] stated that reliability was associated with the volume of the structure, the ratio of volume to surface area for the structure, the magnitude of the interscan interval, and the method of segmentation.

The validation study by Kempton et al. [80] carried on to assess the reproducibility of their segmentation algorithms and FreeSurfer in the same subjects using the same MRI scanner and pulse sequence on publicly available database, Open Access Series of Imaging Studies (OASIS, `www.oasis-brains.org`). The study concluded that both algorithm demonstrated high accuracy and good reproducibility but limitation was observed in segmenting ventricular volume in patients with Alzheimer's disease or healthy subjects with large ventricles.

6.3 Method

Our open source software, *BRAINSCut*, encodes a family of machine-learning techniques in the unified framework efficiently and flexibly. All the data has been processed by using identical procedures, including identical parameter set of pre- and post-processing steps by using cluster computing resources provided at University of Iowa. In section 6.3.1, we points out the importance of the software robustness for the large-scale data analysis and introduce a set of software tools and environments employed in this work. In Section 6.3.2, the report continues to the general flow of the procedure.

6.3.1 Processing Large-Scale MRIs with our Proposed Tool

The robust tool development for large-scale and multicenter data processing requires more than algorithmic procedural robustness. Developers also have to ensure the software robustness across platforms, availability of infrastructures to execute in timely manner, and repeatable evaluation procedure against large amount of data for the cyclic development process. *BRAINSCut* software that we developed in this study is putting continuous endeavor to meet those requirements by investigating 1) software testing for multi-platform and environments, 2) high-performance and high-throughput computing resources, and 3) NiPype, collaborative platform generating language for neuroimaging software development [56].

Software Testing for Multi-Platform All the software developed are tested nightly for multiple platforms including Mac OS-X and Linux by using *CDash*.

> *CDash is an open source, web-based software testing server, which aggregates, analyzes and displays the results of software testing processes submitted from clients located around the world. Developers depend on CDash to convey the state of a software system, and to continually improve its quality.*
>
> *(*excerpt from `http://www.cdash.org`*)*

HPC/HTC High performance and high throughput computing resources at University of Iowa were extensively used to accomplish this research, to test and analyze large-scale multi-center MR data. HPC/HTC are a recently emerging paradigm and defined as following:

HPC (High Performance Computing) systems enable users to run a single instance of parallel software over many processors.

HTC (High-Throughput Computing) serial systems are more suited to running multiple independent instances of software on multiple processors at the same time.

There are **over 3000 multi-modal MRI sessions** to be processed for analysis and the processing takes **about 12-18 hours** per session data with 4-core machine. The time suggested here is minimally calculated because however well planned, in general, the quality of any software evolves through out iterative development process between algorithmic advances and trials/evaluations.

HPC/HTC resources provide a wide variety of software and computing systems that are highly parallel shared memory systems and distributed memory systems (clusters). The paradigm of HTC, the use of many computing resources to do repeatedly similar task, allows us to search for robust methodologies in terms of multiple algorithms and feature groups in relatively short period time. In addition, HPC/HTC made it possible for us to try larger testing cases effectively so that all the unexpected cases can be visited for algorithmic improvements.

NiPype NiPype is *'an open-source, community-developed initiative under the umbrella of NiPy, is a Python project that provides a uniform interface to existing* *neuroimaging software and facilitates interaction between these packages within a single workflow* [56]

Neuroimaging in Python Pipelines and Interfaces is a processing framework that easily creates an unified *workflow* for automating pipelines from multiple software components for testing and development [56]. HPC/HTC resources allow us to process a large amount of data simultaneously in a short period time, which means that it also requires methods to control and examine the large number of processed data more effectively.

All the repeated experimental trials as well as full framework including pre- and post-processing of segmentation are all constructed within NiPype workflow for effective testing and deployment. The pipeline and software packages are publicly available at `https://github.com/BRAINSia/BRAINSTools`.

The automated pipeline of our segmentation framework is summarised in Figure 6.1 and the resulting Nypipe workflow for results analysis is shown in Figure 6.2. These two coupled workflows are constructed to automatically configure necessary inputs, and conduct experiments based on the configuration file, and generate a brief report about the experiments. This automation of experiments greatly shorten the development process for finding optimal parameters of our automated segmentation framework.

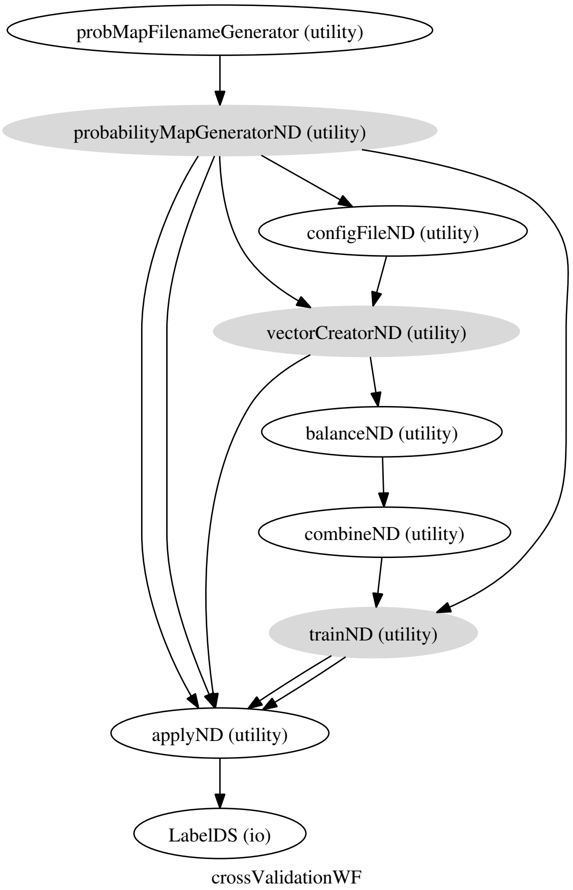


Figure 6.1: NiPype *workflow* is designed to augment necessary software components to conduct *cross-validation*. The workflow graph generated automatically based on the NiPype script. The workflow is designed to take in a single configuration file, which describes all the necessary inputs and parameter arguments. This minimizes operator's intervention effectively. This cross-validation workflow contains processes of region-specific prior generation (*probMapGeneratorND*, ND for node) input vector creation including balancing the number of vectors between classes (ROIs) (*vectorCreatorND*, *balanceND*, and *combineND*), training (*trainND*) from testing data set and applying (*applyND*) on the hold-out data set for k-fold validation. The automation of the testing process is required to effectively employing HPC/HTC resources. Detailed graph including specific I/O is also given in Appendix Figure A.7

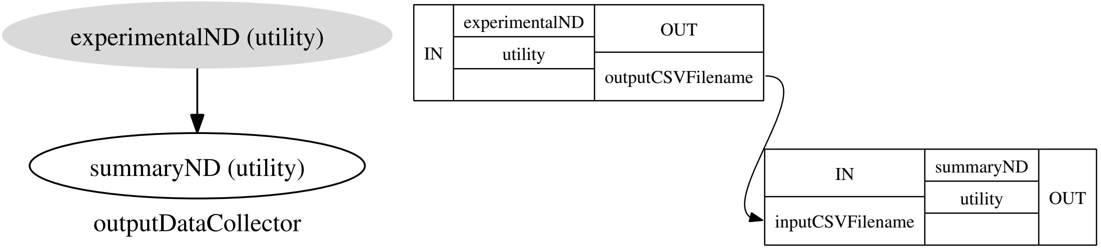


Figure 6.2: Nipype *workflow*, which automatically augments necessary software components and conducts analysis on the results generated from the pipeline presented in Figure 6.1. Left shows the brief workflow and right graph shows the connection of input and output from *experimentalND* to *summaryND*. These two automations of segmentation framework and analysis phase have greatly reduced developers' interactions for the development while highly utilizing HPC/HTC resources by providing automatically generated result analysis documents and charts.

6.3.2 Image Processing

Acquired scans are processed through a fully automated procedure, *BRAINS Auto Workup (BAW)*, improved with SyN registration from the Advanced Normalization Toolkit in the BRAINSTools package. Note that the proposed segmentation algorithm, *BRAINSCut* is now an integral part of the *BAW*. All the scans begin with visual inspection of the raw data so that only images of sufficient quality are subjected to further processing. Each dataset, T1- and/or T2-weighted images, are processed together to improve the robustness of the procedure from complimentary information provided by multiple modalities and repeated scans. The best-rated T1-weighted image is spatially normalized based on prominent landmarks in MRI, including anterior (AC) and posterior commissure (PC), and mid-sagittal plane. The remaining scans acquired of the same session are then rigidly aligned to the AC-PC aligned T1 image, and simultaneously processed by the automated bias-field correction (ABC) algorithm, BRAINSABC. For each given modality, BRAINSABC produces an average of independently bias-field corrected MR images resampled in 1mm×1mm×1mm and their respective corresponding 17 tissue probability maps, including white matter, grey matter, and CSF. At this point, all longitudinal scan sessions for given subjects are used jointly to build a subject specific atlas that best represents the average longitudinal shape with respect to minimum mean square error of displacement. This joint session template building is a normalizing step that uses the all scan sessions for a given subject to maximize consistency of subsequent measurements across scanner variation inherent in long-running longitudinal studies. The resulting data set of bias-corrected average T1 and/or T2 images are subsequently segmented for subcortical structures using an automated segmentation framework, BRAINSCut. BRAINSCut employs robust random forest machine learning that has been validated on multi-site MR data. The subcortical structures of interest include nucleus accumben, caudate nucleus, putamen, hippocampus, and thalamus. The result of this procedure were again visually inspected and resulted in a success rate greater than 90%. All the development processing was blinded to clinical data, such as HD gene-expansion status, gender, and age.

6.3.3 Quality Assessment in terms of validity and reliability:

Our study provides evaluation of both validity and reliability of the segmentation framework. As we briefly mentioned in the introduction of this study (Section 6.1), there are two main criteria in determining the quality of measuring instrument/software : 1) validity and 2) reliability [85]. As we reviewed in the Section 6.2, a number of validation studies among different segmentation algorithms have already been conducted. Most of studies, however, only focused on the reliability but validity. We believe that providing segmentation validation study with regard to both validity and reliability advances the understanding of current status of automated segmentation framework, and so allows consequence quality studies derived from the automated measures. In that respect, we evaluated our segmentation framework with regard to those two indicators of validation: 1) validity and 2) reliability.

> *Validity is often defined as the extent to which an instrument measures what it purports to measure. Validity requires that an instrument is reliable, but an instrument can be reliable without being valid. ······ Validity is the extent to which the interpretations of the results of a test are warranted, which depend on the test's intended use (i.e., measurement of the underlying construct). ······ Because there is no statistical test to determine whether a measure adequately covers a content area or adequately represents a construct, content validity usually depends on the judgment of experts in the field.* *(*excerpted from [85] *)*

6.4 Multicenter Reliability Assessment through Traveling Human Phantom Data

To assess multicenter reliability, the automated segmentation tool is applied eight independent subcortical structure segmentations of five subjects taken from eight sites. The evaluation data from this study is described in Section 6.4.1 and the segmentation results with related discussions are given in Section 6.4.2.

6.4.1 Data

We have utilized traveling human phantom (THP) data for validation of our proposed tool. THP data consists of five subject scanned at eight sites repeatedly over a month period. Note that THP data was originally planned and collected to evaluate diffusion tensor imaging (DTI) process as reported in [94]. Eight sites participated in this multicenter image collection consists of two MR vendors of distinguished imaging histories: Siemens and Phillips. The sites involved in this study had either a Siemens 3T TIM Trio scanner (gradient strength $=45mT/m$, slew rate $= 200\ T/m/sec$) or Philips 3T Achieva scanner (gradient strength $= 80mT/m$, slew rate $=200T/m/sec$). Five healthy control subjects were recruited into this multicenter imaging study after informed consent was obtained in accordance with the Institutional Review Board at each of the imaging sites. All five subjects were imaged at the eight sites within a 30-day period. Collected data includes T1- and T2-weighted multi-modal MR images, acquired using using three-dimensional (3D) T1 weighted (MP-RAGE) and T2 (SPACE) sequences at each center.

Each MRI anatomical volume was processed with the standard BAW procedure 6.3.2. After visual inspection stage based on our standard protocol, seven scan sessions are removed from further analysis due to the low quality of T1 images (Marked as **(X)** in Table 6.1). The common reason of low score was a insufficient coverage of whole brain region as shown in Figure 6.4, which, in turn, results in failure of spatial normalization of BAW process.

Table 6.1: THP Data Quality Report

Center	*Vendor*	*Visual Inspection Scores [T1s(repeat)/T2s]*				
		THP 1	*THP 2*	*THP 3*	*THP 4*	*THP 5*
CCF	Siemens	9/10	9/10	10/8	8/10	9/10
IOWA	Siemens	9,4/8	10/10	8/8	8/6	6/10
MGH	Siemens	10/7	10/10	8/8	8/9	10/8
UCI	Siemens	10/9	8/10	9/10	8/8	9/8
UMN	Siemens	7/7	8/8	10/9	10/8	8/8
DART	Philips	10(2),0/10,0	10(5)/8	8/8	10(2),0/10	10(3),8,8/9
UW	Philips	0/8 **(X)**	8/8	0/8 **(X)**	0/10 **(X)**	8/10
JHU	Philips	0,0/8**(X)**	10,8/8	0,0/10 **(X)**	0,0/8 **(X)**	0,0/5 **(X)**

Quality report of THP data from the experts' visual inspection. For each scans of a scan session, visual inspection score ranges from one to ten for the worst (1) and the best (10) image qualities. Each MR session can have multiple scans including more than one T1- and T2-weighted images. Score S is reported T1 first and followed by T2 separated by slash(/). If there is multiple scans that rated identical, the number of scans n are reported in parenthesis: [S_{T1_1}, S_{T1_2} (n),... / S_{T2_1}, S_{T2_2} (n),..], where S_I is a visual inspection score for scan I. The only data rated above > 5 is proceeded to the standard BAW procedure and acquired six subcortical structure segmentation results. Note that eight scans from UW and JHU are excluded from processing because no T1-weighted image is remained after the quality control.

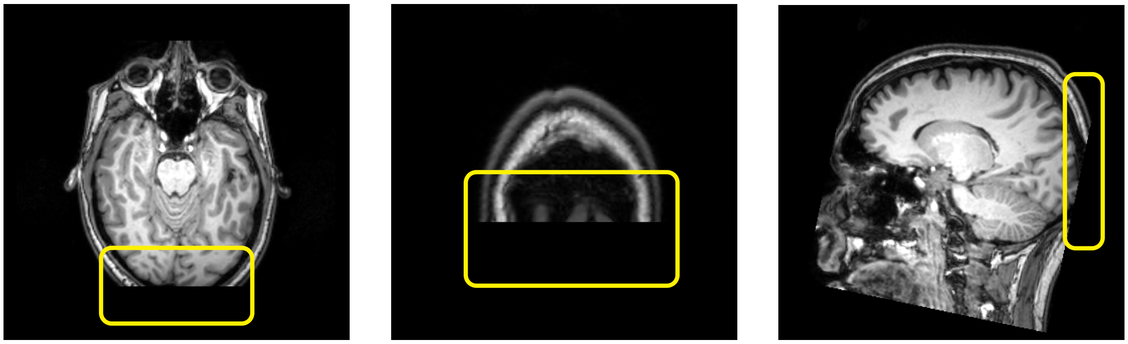


Figure 6.3: Traveling Human Phantom Low Quality Scored MR Image Example: A Scan have an insufficient head region coverage to the posterior of brain. This insufficiency leads to the failure at the pre-processing stage while taking the scan into common AC-PC aligned space as described in Section 1.6.1.

6.4.2 Results and Discussion

Inter- and intra-center reliability are assessed through the THP results for six subcortical structures. We focus on the automatic segmentation reliability between repeated measures of identical subject. Inter- and intra-center reliabilities are measured in terms of coefficient of variation (CV). CV represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one set of data to another (`http://www.investopedia.com/terms/c/coefficientofvariation.asp`) regardless of absolute measurement unit:

$$CV = \frac{\sigma}{\mu} \times 100\% \tag{6.1}$$

Because the variance σ and the mean μ share the same units of measurements, the units cancel out and leaves CV a *dimensionless* number [110]. In practice, since true population mean μ and variance σ is usually unknown, CV is typically estimated with standard deviation s and sample mean $\bar{x}$:

$$\widehat{CV} = \frac{s}{\bar{x}} \times 100\% \tag{6.2}$$

For multi-site study, the lower CV is desired as for small variation between sites. One possible disadvantage of CV is when the mean is close to zero, the CV will approach infinity and thus sensitive to small change in the mean. Keeping in mind its sensitivity to the mean fluctuation in the case of small mean data, CV is mainly used to assess inter- and intra-site reliability.

6.4.2.1 Volume Correspondence

Drived subcortical volumes from our proposed method displays small variations between repeated measures of same subject across site as shown in Figure 6.4. As noted in Table 6.1, those scans with the low visual inspection score, therefore poor scan quality, have been omitted from the further analysis. We can also recognize more fluctuation of measured volumes on larger structures, such as thalamus and putamen while it is less obvious in smaller structures such as nucleus accumben and globus pallidus. Volume mean and standard deviation of six subcortical structures in both hemispheres across five subjects are also summarized in Table 6.2. For a subject, the CV of the measured volumes ($CV = sd/mean\%$) from repeated scans at multicenter are reported as well. In a corresponding manner to the results in Figure 6.4, the smaller the average volume is, the more error (variation) the data presents. The overall variation among measured volumes were in range of $3 \sim 10\%$, and the deviation were less with structures with larger volumes.

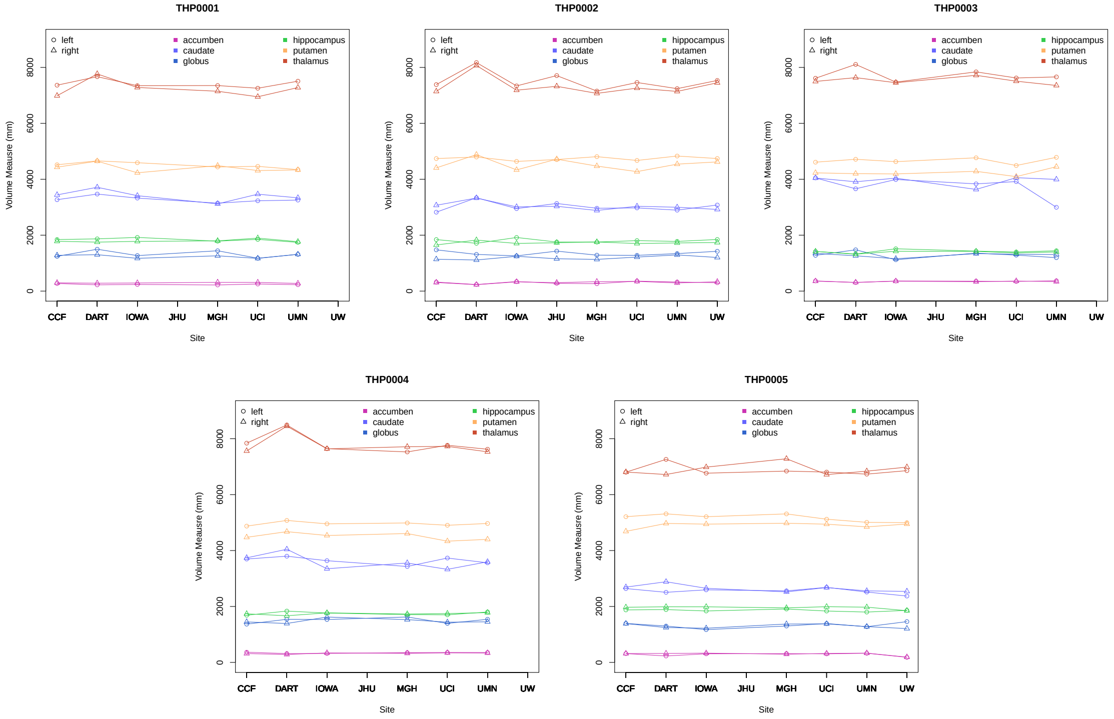


Figure 6.4: For Traveling Human Phantom study, five healthy subjects were scanned at eight sites in a month and six subcortical volumes are measured by using our proposed segmentation method. Seven scan sessions are eliminated from our analysis due to the low quality of T1 images (Marked as **(X)** in Table 6.1). All the process successfully identified six subcortical structures other than those which failed at the visual inspection stage. Intrasubject volumetric differences, however, is observed to some extent as we analyze and discuss later in this section.

Table 6.2: Traveling Human Phantom Data Mean and Standard deviation (sd) measured from MRI by using our proposed approaches.

ROI		TPH01 (n=6)	TPH02 (n=8)	TPH03 (n=6)	TPH04 (n=6)	TPH05 (n=7)	Mean
		mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	
accumben	L	238.8 (18.5)	302.8 (38.8)	349.8 (21)	342.3 (18.6)	282.9 (50.9)	303.32 (29.56)
		8%	13%	6%	5%	18%	10%
	R	294 (13.6)	309.4 (35.2)	342.3 (15.5)	322.8 (19.8)	300.3 (47)	313.76 (26.22)
		5%	11%	5%	6%	16%	8%
caudate	L	3282.5 (99.7)	3019.4 (150.9)	3741.2 (355.2)	3642.7 (120.3)	2553.3 (94.6)	3247.82 (164.14)
		3%	5%	9%	3%	4%	5%
	R	3414.3 (175.8)	3034.8 (124.8)	3946.5 (147.2)	3602.5 (243.9)	2645.9 (115.9)	3328.8 (161.52)
		5%	4%	4%	7%	4%	5%
globus	L	1323.2 (112.5)	1350.3 (76.2)	1285.8 (112.9)	1500.5 (89.8)	1327.1 (87.6)	1357.38 (95.8)
		9%	6%	9%	6%	7%	7%
	R	1250.2 (58.9)	1187.1 (58.6)	1289 (65.1)	1480 (75.4)	1299 (72.1)	1301.06 (66.02)
		5%	5%	5%	5%	6%	5%
hippocampus	L	1834.5 (56.9)	1800.4 (61.8)	1421.3 (55.3)	1749.8 (55.2)	1858.9 (34.2)	1732.98 (52.68)
		3%	3%	4%	3%	2%	3%
	R	1794.8 (46.9)	1728 (46.6)	1390.7 (37.9)	1739 (36.3)	1959.3 (46.7)	1722.36 (42.88)
		3%	3%	3%	2%	2%	3%
putamen	L	4501.8 (102.3)	4741.1 (63.2)	4665.8 (101.3)	4960.8 (64.6)	5167.3 (120.8)	4807.36 (90.44)
		2%	1%	2%	1%	2%	2%
	R	4406.8 (135.6)	4530.5 (190)	4240.7 (108.5)	4505.3 (116.6)	4902.3 (96.8)	4517.12 (129.5)
		3%	4%	3%	3%	2%	3%
thalamus	L	7415.3 (135.4)	7498.8 (301.1)	7719.2 (203.9)	7815.5 (319)	6867.1 (165.5)	7463.18 (224.98)
		2%	4%	3%	4%	2%	3%
	R	7235.2 (269.7)	7330.1 (300.6)	7527.2 (118.5)	7769.8 (312)	6904.7 (184.7)	7353.4 (237.1)
		4%	4%	2%	4%	3%	3%

For five subjects, measured volumes of six subcortical structures are shown as well as CVs of each subjects.

The volume difference between site was formally tested by the analysis of variance (ANOVA) as shown in Table **??**. We tested if measured means across subjects are different between eight sites:

H_0 : Volume means of five subjects are same across eight different sites.

To employ ANOVA, we first confirmed that our data meet a required assumption, homoscedasticity [1] and then tested the measure differences across site. Our statistical test suggests that there is no significant differences between eight sites in measuring subcortical volumes (Table **??**).

Sample size is also determined based on the mean of mean volume and standard deviation of subjects across sites reported in in Table 6.2 at the last column. This measurement is important for designing efficient clinical and research trials. The required sample size to detect 5% and 10% mean volume difference between two groups are shown in Figure 6.5. We varied the range of power from 0.5 to 1.0 on each computation. The two-sample t-test formula is used to calculate required sample sizes along the power levels assuming balanced but unpaired with equal variance design.[2] To detect 5% and 10% mean volume difference with about 0.8 power, appropriate sample size would be 30 and 120 for nucleus accumben respectively.

[1] *Homoscedasticity*: The term is also known as homogeneity of variance in statistics

[2] A free software programming language and a software environment for statistical computing *R* package *'samplesize'* is employed for the computation. The package is available at `http://www.inside-r.org/packages/cran/samplesize/docs/n.ttest`

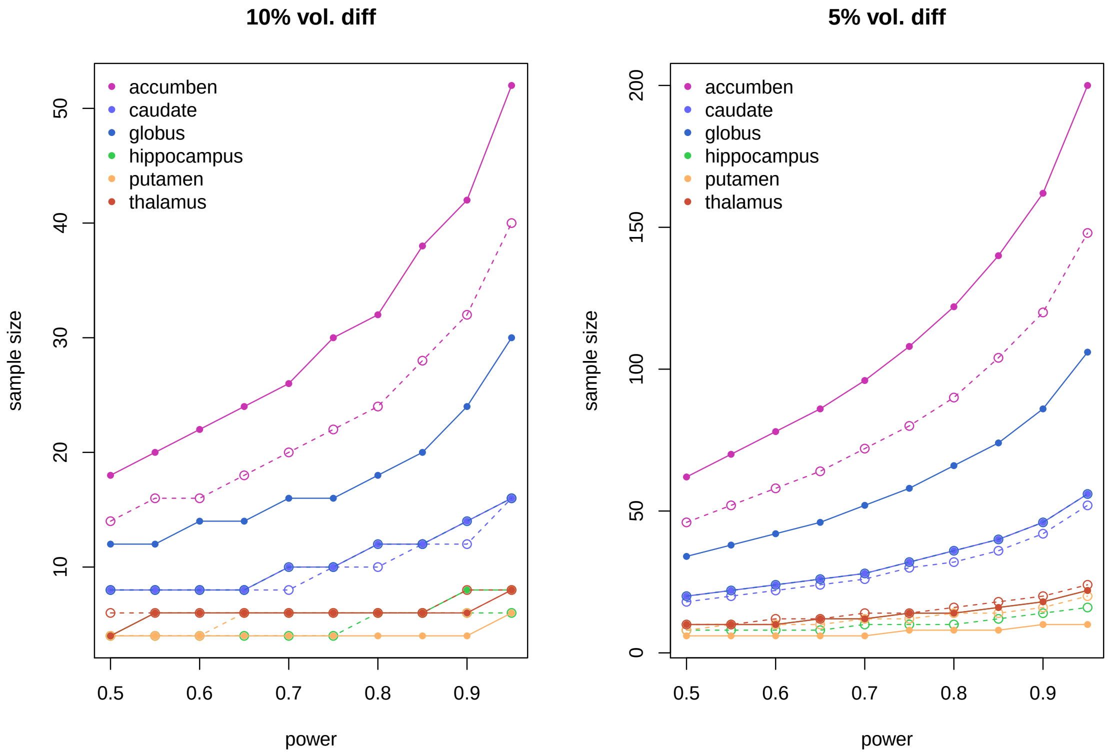


Figure 6.5: The estimated required sample size to detect 5% (left) and 10% (right) volume changes for the corresponding subcortical structures: 1) Nucleus accumben (accumben), 2) Caudate, 3) Globus Pallidum (globus), 4) Hippocampus, 5) Putamen, and 6) Thalamus. Solid line and dashed line represents structures located in left and right hemisphere, respectively. Mean and standard deviation of each volume is estimated from THP data by averaging over all the subject's mean and standard deviation from eight sites (see the last column of Table 6.2).

6.4.2.2 Inter- and Intra-center reliability

The inter- and intra-center reliability is evaluated by employing CVs of measured volumes. CVs for *intracenter reliability* is computed from a subject that acquired multiple sessions with same scanner, same protocol at the same site and reported for all six subcortical structures. CVs for *intercenter reliability* is computed from each subjects that scanned across eight sites acquired within a month. Intra and inter-center reliability from the CVs are shown in Figure 6.6, and clearly depicts the better intracenter reliability than intercenter reliability. Both intercenter and intracenter CVs were reasonably low $(< 15\%)$ while demonstrating better intracenter reliability in general. In addition, association between general size of the structure and the CV can also be observed here: the bigger size the structure is, the smaller the CV is. For instance, nucleus accumben result in a greater intercenter variability than intracenter variability.

6.5 Repeated-Measure Reliability

This section provides a result and analysis of repeated in-vivo MRIs to provide a scan-rescan (test-retest) reliability. Repeated-measure reliability evaluate the stability of measures and internal consistency of measurement instruments [85]. Specifically, our research interest is to identify disease progression from the automated morphometric analysis therefore measurement stability across disease progression is crucial. In this respect, we have chosen the CAP score [92] of HD study [180] as an indicator of HD disease status; the CAP group reflects the individual's progression through the

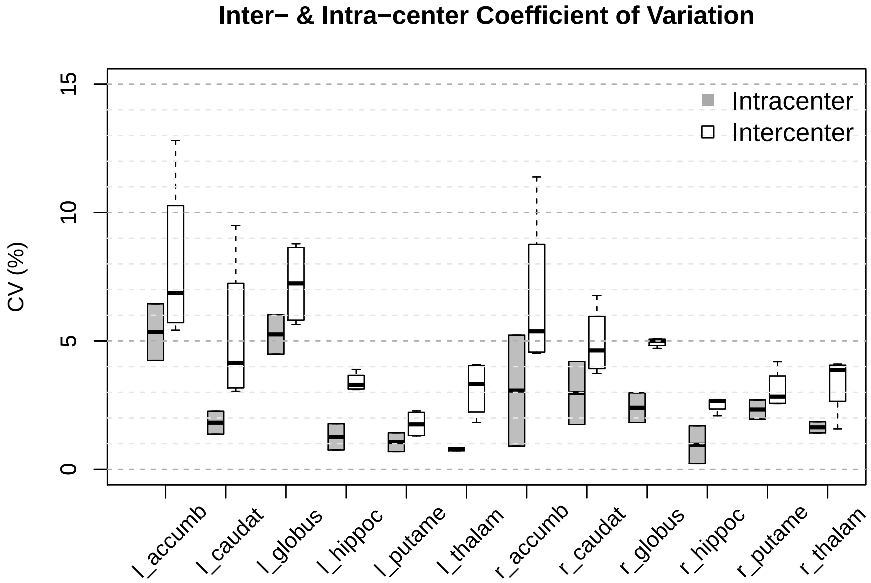


Figure 6.6: Both intercenter and intracenter coefficient of variation (CV), as measures of reliability measure of intra- and inter-center, are reasonably low CV ($< 15\%$) while better reliability of intracenter (lower intracenter CV) than intercenter is clearly demonstrated in this figure. CV is computed from Traveling Human Phantom (THP) volumetric measures of six subcortical structures in both left and right hemisphere. THP study conducted for five human subjects scanned at all eight multiple centers in a month.

HD disease process, from presymptomatic through manifest HD, based on CAG and age. It is meant to encompass terms such as "disease burden" and "genetic burden" that have been used in previous literature. CAP group is determined by the scaled CAP score, CAP_S, with two cutoff of $CAP_S = 0.67$ (lower cutoff) and $CAP_S = 0.85$ (upper cutoff). The formula for CAP_S is as follows:

$$CAP_S = Age_0 \times (CAG - 33.6600)$$

where Age_0 represents age of the participant at the time of scan for this study (i.e., baseline). Details and justification about CAP_S and CAP group are fully described in [180].

Note that particular evaluation is a retrospective analysis, where the evaluation data is handpicked from the entire PREDICT-HD data after processing completed. The section (Section 6.5.1) describes how the evaluation data were selected in details and Section 6.5.2) presents results and discuss the repeated-measure reliability of our proposed method.

6.5.1 Data

From Predict-HD [127] cohort, we have identified 287 repeated scans acquired less than a week (< 8 days), where insignificant morphological brain change is expected. All the scans are acquired for T1- and T2-weighted MRIs at the same sites within eight days. Of 287 paired data, all four groups of interest are included: a) control group ($n = 102$), b) low CAP group ($n = 70$), c) med CAP group ($n = 71$), and d) high CAP group ($n = 26$).

Again, since the identified scans for this evaluation were not intended for the test-retest analysis at a time of data collection, and thus scanning protocols as well as any other relative conditions may vary. In this study, we conduct analysis assuming that this reptrospective rescanned data provides more realistic estimates about applications to clinical trials settings of MRI collection.

Table 6.3: Scan-Rescan Reliability Test Data Demographic

Gender	Control	Low	Med	High	Total
Female	73	53	64	20	210
Male	35	25	10	7	77
Total	108	78	74	27	287

Demographic of scan-rescan data for reliability assessment. $n = 287$ scans are identified from the entire PREDICT-HD data that we completed the processing in 2013 for three CAP groups (low, med, and high) and normal healthy control All the scan-rescan data that acquired repeatedly less than 7 days ($\leq 7\ days$).

6.5.2 Results and Discussion

The effect of clinical status, HD disease progression encoded by CAP group, on the repeated-measure reliability is investigated and discussed. Volume correspondence among automated measures of six subcortical structures from repeated MRI scans are reported and discussed in Section 6.5.2.2. In addition, a formal statistical model is tested in Section 6.5.2.3 to see the effect of the disease progression on the method reliability.

6.5.2.1 Intraclass correlation coefficient (ICC)

Intraclass correlation coefficient is often used in reliability studies where n targets are rated by k judges [145]. We have utilized two ICC versions for assessing agreement $ICC(A)$ and consistency $ICC(C)$:

$$ICC(A) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_e^2} \tag{6.3}$$

$$ICC(C) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}, \tag{6.4}$$

where σ_s is within subject standard variation that only caused by biological variation between tests and σe represents variation caused by measurement error. For accuracy results, the ICCs quantifies how well the automated segmentations agree with the gold standard, for reproducibility measures the $ICC(C)$ values quantifies the consistency of the segmentations [80]. For test-retest data, acquired under similar condition in a short period time, true within subject differences σ_s, biological brain volume changes, should be small and the reliable method will yield the high ICC.

6.5.2.2 Volume Correspondence

Mean and standard deviation of six subcortical volumes are summarized in Table **??**. Absolute volume differences are reported in Table **??** to characterize random error presented in the volumetric measure that may confound the meaningful biological volumetric changes. The mean volume differences between baseline and follow-up scans as generally quite small ($< 5\%$) except for globus pallidus in high

CAP group (-14.16% and -12.81% for left and right globus pallidus). Correlation between paired scans is also generally high except for nucleus accumben and globus pallidus. Because nucleus accumben and globus pallidus are rather small in size and their boundaries are hard to define, it should not be surprising that less reliability were observed than other structures. The greatest single structural volume increase was 44.4% with nucleus accumben where as the greatest single volume decrease was -95.3% with globus pallidus.

ICC values between baseline and follow-up scans in the scan-rescan data are also provided in Figure 6.7, demonstrating degree of variation according to structures of interest. ICC evaluation is also consistent to the absolute volume differences, where low reliability for nucleus accumben and globus pallidus were observed. Note that $ICC(A)$ and $ICC(C)$ are appeared to be similar each other, which supports our claims of consistency as well.

6.5.2.3 Sensitivity: Effect of Disease Status on Volume Measure

Sensitivity of volume measurement against HD disease status is investigated and discussed. The results are reported in Table 6.5 and Table 6.6 for structures in left and right hemisphere, respectively. Only the caudate nucleus had statistical significant relationship with high CAP group.

Sensitivity of volume measurement against clinical variable performed as one of reliability assessments of the proposed segmentation results. From the repeated scans, the sensitivity (or relation) of measured volume to clinical variables are investigated. Among various independent clinical variables, we specifically interested in the effect of degree of disease progression on the volume measurement. That is, if the scan-rescan reliability is independent on the status of disease progression, so that the measured results can be used to answer the research questions of interest, such as how the volume of ROI changes or is different according to the disease status.

We conducted ANOVA analysis, as an extension of two-sample t-test for comparison between four groups including healthy control, low, mid, and far CAP groups.

$$|Vol_{time1} - Vol_{time2}| = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 CAP, \tag{6.5}$$

where $\beta_3 = I_{low}\beta_{(3,low)} + I_{mid}\beta_{(3,mid)} + I_{high}\beta_{(3,high)}$ with I is indicator variable for CAP groups. Our main interest is *if there is differences between* $\beta_{(3,\cdot)}$.

Table 6.4: Absolute volume differences on the scan-rescan (test-retest) data between baseline (B) and follow-up (F) volumetric measures from *BRAINSCut*.

ROI		*B.mean*	*B.sd*	*F.mean*	*F.sd*	*D.mean*	*D.sd*	*D.%*	*D.ρ*
accum [C]	L	277.1	54	273	43.4	4.13	59.72	1.33	0.26
accum [L]	L	251.8	50.3	257.1	46.7	−5.23	55.47	−1.98	0.35
accum [M]	L	242.7	43.2	230.8	40.4	11.93	48.34	5.08	0.33
accum [H]	L	202	38.2	216.4	56.5	−14.38	40.83	−6.95	0.69
accum [C]	R	292.3	48.9	291.8	48.7	0.5	25.42	0.15	0.86
accum [L]	R	267.2	41.2	270.4	40.1	−3.2	26.59	−1.35	0.79
accum [M]	R	255.2	42	257.5	38.9	−2.28	23.64	−1.04	0.83
accum [H]	R	214.3	39.6	224	42.7	−9.69	19.9	−4.64	0.89
caud [C]	L	3,165.4	525.8	3,169	510.7	−3.57	348.75	−0.13	0.77
caud [L]	L	3,081.8	472.6	3,083.9	497.5	−2.07	314.59	−0.08	0.79
caud [M]	L	2,784.4	535.1	2,812.3	566.9	−27.99	347.24	−1.49	0.8
caud [H]	L	2,197.1	464.7	2,083.1	471.3	114.04	402.07	5.32	0.63
caud [C]	R	3,194.2	538.5	3,236.8	545.8	−42.58	384.14	−1.51	0.75
caud [L]	R	3,079.9	461.3	3,113.9	471.8	−34	371.82	−1.3	0.68
caud [M]	R	2,790.7	535.5	2,863.6	551.1	−72.85	300.72	−4.12	0.85
caud [H]	R	2,156.7	536.5	2,056.4	540.8	100.31	319.07	4.41	0.82
glob [C]	L	1,066.2	204.9	1,049.8	181.3	16.35	257.81	1.43	0.11
glob [L]	L	1,010.1	153.5	974.9	172.3	35.17	199.39	4.46	0.26
glob [M]	L	900.9	193.8	882.5	191	18.42	187.46	2.93	0.53
glob [H]	L	614.4	159.1	683.2	169.9	−68.81	155.03	−14.16	0.56
glob [C]	R	1,061.9	182.3	1,060.1	187.7	1.85	192.86	0.17	0.46
glob [L]	R	993.1	178.3	967	180.6	26.14	140.53	3.33	0.69
glob [M]	R	870.4	173.9	853.4	165.8	16.99	121.5	3.13	0.75
glob [H]	R	592.8	132.9	650.8	142.2	−58.04	129.11	−12.81	0.56
hipp [C]	L	1,683.3	196.5	1,696.2	188.9	−12.95	138.65	−0.77	0.74
hipp [L]	L	1,646.5	279.3	1,634.8	287.3	11.7	139.87	0.57	0.88
hipp [M]	L	1,662.5	174.7	1,687.7	201.2	−25.13	133.52	−1.78	0.76
hipp [H]	L	1,556.2	204.3	1,554.4	170.5	1.73	137.44	0.1	0.75
hipp [C]	R	1,601.8	223.1	1,578.9	202.8	22.9	154.99	1.38	0.74
hipp [L]	R	1,542.6	266.9	1,510.1	256.6	32.43	153.13	1.72	0.83
hipp [M]	R	1,571.8	198.5	1,556.7	181.1	15.06	147.06	1.13	0.7
hipp [H]	R	1,444	148.8	1,481.6	152.1	−37.65	162.03	−2.61	0.42
puta [C]	L	4,255.1	577.9	4,244.9	548.2	10.24	244.59	0.26	0.91
puta [L]	L	3,982.7	520.6	3,983.9	515	−1.16	237.2	−0.03	0.9
puta [M]	L	3,558.2	522	3,587.4	539.2	−29.25	197.51	−1.13	0.93
puta [H]	L	2,841.3	560	2,814.4	530.4	26.88	185.96	1.02	0.94
puta [C]	R	4,039.5	589.1	4,073.9	563.3	−34.33	204.33	−0.89	0.94
puta [L]	R	3,804.1	504.3	3,777.8	534.9	26.33	252.11	0.82	0.88
puta [M]	R	3,391.8	521.2	3,410.2	518.5	−18.38	153.48	−0.82	0.96
puta [H]	R	2,706	543.3	2,664.7	514.1	41.23	144.23	1.69	0.96
thal [C]	L	6,912.1	805.1	6,903.6	743.9	8.47	298.91	0.14	0.93
thal [L]	L	6,836.5	704	6,790.3	678.3	46.23	262.09	0.57	0.93
thal [M]	L	6,797.2	682.1	6,764.8	569.5	32.39	431.43	0.5	0.78
thal [H]	L	6,546.2	478.4	6,642.1	656.9	−95.85	440.66	−1.4	0.74
thal [C]	R	6,865.2	864.5	6,825.1	797.5	40.14	244.57	0.63	0.96
thal [L]	R	6,715.7	729.5	6,685.6	699.2	30.07	238.67	0.37	0.95
thal [M]	R	6,716.1	689.1	6,651.2	620.7	64.96	379.54	1.04	0.84
thal [H]	R	6,547.5	589.5	6,654	825.1	−106.42	458.71	−1.6	0.84

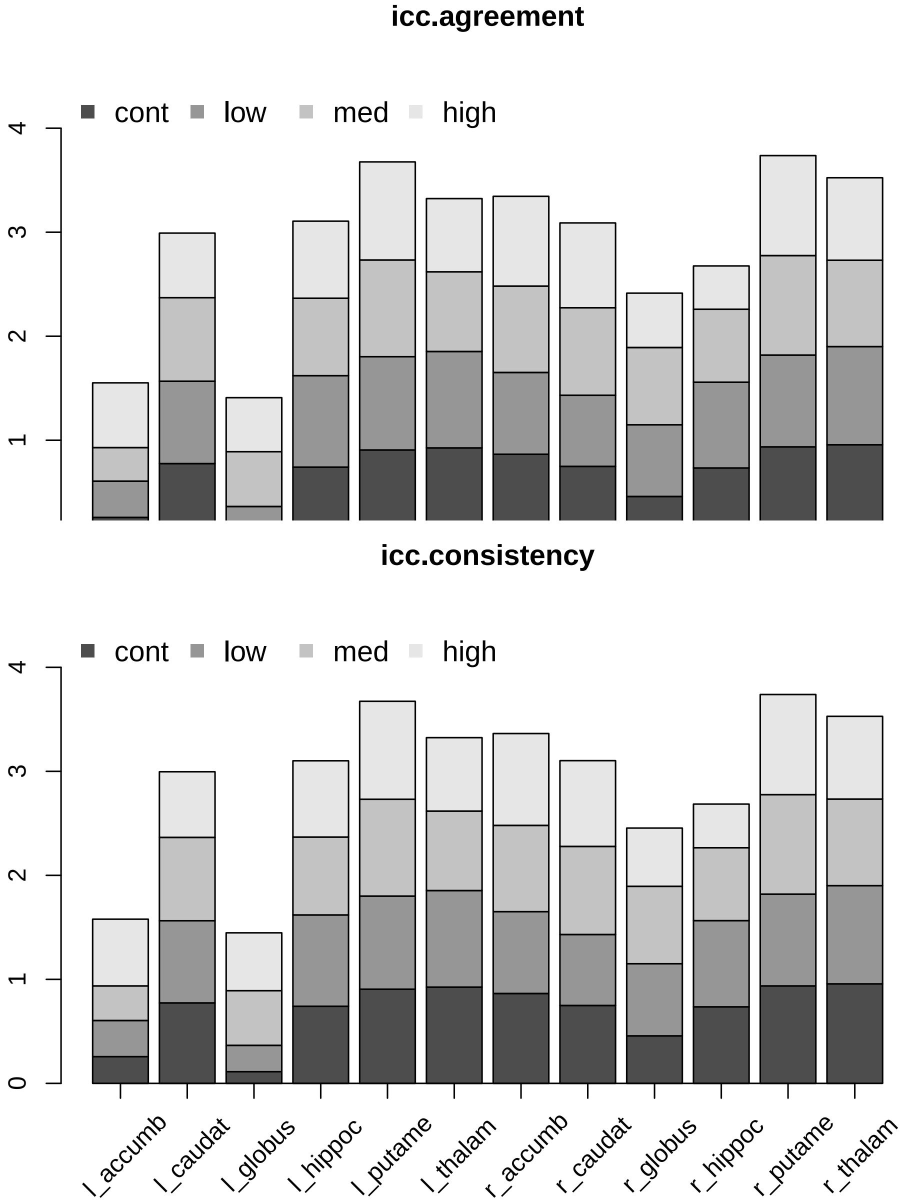


Figure 6.7: ICCs across CAP group stacked and demonstrates reliability variation for each ROIs including nucleus accumben, caudate nucleus, globus pallidus, hippocampus, putamen, and thalamus. Variation across CAP groups (box size variation between different grey levels on a stacked bar), however, is not noticeable from the graph. Note that CAP group effect on automated measurement reliability were not statistically significant except for left caudate nucleus with high CAP group at $\alpha = 0.001$ level (Table 6.5 and 6.6)

Table 6.5: ANOVA table of automated segmentation reliability study scan-rescan in-vivo MRI of six subcortical structures in left hemisphere.

ROI	*coeff.*	*estimate*	*std*	*t-value*	*P-value*	*signif.level*
(l) accumbens	(Intercept)	13.536	2.158	6.27	$1.46{\cdot}10^{-9}$	***
	scan.age	0.005	0.042	0.11	$9.13{\cdot}10^{-1}$	
	low	−0.183	1.285	−0.14	$8.87{\cdot}10^{-1}$	
	med	−0.462	1.164	−0.4	$6.92{\cdot}10^{-1}$	
	high	−1.082	1.614	−0.67	$5.03{\cdot}10^{-1}$	
	gender.male	−0.467	1.077	−0.43	$6.65{\cdot}10^{-1}$	
(l) caudate	(Intercept)	5.962	1.397	4.27	$2.75{\cdot}10^{-5}$	***
	scan.age	0.017	0.027	0.64	$5.20{\cdot}10^{-1}$	
	low	−0.161	0.831	−0.19	$8.46{\cdot}10^{-1}$	
	med	0.843	0.753	1.12	$2.64{\cdot}10^{-1}$	
	high	4.793	1.044	4.59	$6.91{\cdot}10^{-6}$	***
	gender.male	0.485	0.697	0.7	$4.87{\cdot}10^{-1}$	
(l) globus	(Intercept)	13.281	1.984	6.7	$1.29{\cdot}10^{-10}$	***
	scan.age	0.048	0.038	1.24	$2.15{\cdot}10^{-1}$	
	low	−1.501	1.18	−1.27	$2.05{\cdot}10^{-1}$	
	med	−1.185	1.07	−1.11	$2.69{\cdot}10^{-1}$	
	high	1.188	1.483	0.8	$4.24{\cdot}10^{-1}$	
	gender.male	−1.313	0.99	−1.33	$1.86{\cdot}10^{-1}$	
(l) hippocampus	(Intercept)	0.461	1.082	0.43	$6.70{\cdot}10^{-1}$	
	scan.age	0.087	0.021	4.14	$4.73{\cdot}10^{-5}$	***
	low	1.329	0.644	2.06	$4.00{\cdot}10^{-2}$	.
	med	0.182	0.584	0.31	$7.56{\cdot}10^{-1}$	
	high	−0.396	0.809	−0.49	$6.25{\cdot}10^{-1}$	
	gender.male	−0.437	0.54	−0.81	$4.19{\cdot}10^{-1}$	
(l) putamen	(Intercept)	2.382	0.779	3.06	$2.47{\cdot}10^{-3}$	*
	scan.age	0.015	0.015	0.98	$3.29{\cdot}10^{-1}$	
	low	−0.215	0.464	−0.46	$6.43{\cdot}10^{-1}$	
	med	0.239	0.42	0.57	$5.70{\cdot}10^{-1}$	
	high	0.384	0.583	0.66	$5.10{\cdot}10^{-1}$	
	gender.male	0.42	0.389	1.08	$2.81{\cdot}10^{-1}$	
(l) thalamus	(Intercept)	0.401	0.733	0.55	$5.85{\cdot}10^{-1}$	
	scan.age	0.038	0.014	2.68	$7.85{\cdot}10^{-3}$	.
	low	0.374	0.436	0.86	$3.92{\cdot}10^{-1}$	
	med	0.251	0.395	0.63	$5.26{\cdot}10^{-1}$	
	high	0.319	0.548	0.58	$5.61{\cdot}10^{-1}$	
	gender.male	−0.136	0.366	−0.37	$7.11{\cdot}10^{-1}$	

No effect of disease progression, *CAP* group, on automated method's reliability is detected other than caudate of *high* group: $CV = \beta_0 + \beta_1 AGE + \beta_2 CAP + \beta_3 Gender$. Significant level is marked with [***], [**], [*], and [.] for < 0.001, < 0.01, < 0.05, and < 0.1, respectively.

Table 6.6: ANOVA table of automated segmentation reliability study on scan-rescan in-vivo MRI of six subcortical structures in right hemisphere

ROI	*coeff.*	*estimate*	*std*	*t-value*	*P-value*	*signif.level*
(r) accumbens	(Intercept)	4.937	1.295	3.81	$1.71 \cdot 10^{-4}$	***
	scan.age	−0.001	0.025	−0.06	$9.56 \cdot 10^{-1}$	
	low	0.853	0.771	1.11	$2.69 \cdot 10^{-1}$	
	med	0.509	0.698	0.73	$4.66 \cdot 10^{-1}$	
	high	1.125	0.968	1.16	$2.46 \cdot 10^{-1}$	
	gender.male	−0.341	0.646	−0.53	$5.98 \cdot 10^{-1}$	
(r) caudate	(Intercept)	8.373	1.432	5.85	$1.48 \cdot 10^{-8}$	***
	scan.age	−0.031	0.028	−1.1	$2.71 \cdot 10^{-1}$	
	low	−0.164	0.852	−0.19	$8.48 \cdot 10^{-1}$	
	med	−0.212	0.772	−0.27	$7.84 \cdot 10^{-1}$	
	high	2.804	1.071	2.62	$9.35 \cdot 10^{-3}$	**
	gender.male	0.67	0.715	0.94	$3.49 \cdot 10^{-1}$	
(r) globus	(Intercept)	10.886	1.849	5.89	$1.19 \cdot 10^{-8}$	***
	scan.age	−0.014	0.036	−0.39	$6.94 \cdot 10^{-1}$	
	low	−1.973	1.1	−1.79	$7.41 \cdot 10^{-2}$	.
	med	−1.82	0.997	−1.83	$6.91 \cdot 10^{-2}$	.
	high	4.118	1.383	2.98	$3.16 \cdot 10^{-3}$	*
	gender.male	0.586	0.923	0.63	$5.26 \cdot 10^{-1}$	
(r) hippocampus	(Intercept)	5.663	1.198	4.73	$3.71 \cdot 10^{-6}$	***
	scan.age	0.021	0.023	0.89	$3.74 \cdot 10^{-1}$	
	low	0.44	0.713	0.62	$5.38 \cdot 10^{-1}$	
	med	−0.758	0.646	−1.17	$2.42 \cdot 10^{-1}$	
	high	0.703	0.896	0.78	$4.33 \cdot 10^{-1}$	
	gender.male	−1.946	0.598	−3.26	$1.28 \cdot 10^{-3}$	**
(r) putamen	(Intercept)	1.881	0.81	2.32	$2.09 \cdot 10^{-2}$	*
	scan.age	0.007	0.016	0.46	$6.48 \cdot 10^{-1}$	
	low	0.575	0.482	1.19	$2.34 \cdot 10^{-1}$	
	med	0.165	0.437	0.38	$7.06 \cdot 10^{-1}$	
	high	0.866	0.605	1.43	$1.54 \cdot 10^{-1}$	
	gender.male	1.335	0.404	3.3	$1.09 \cdot 10^{-3}$	*
(r) thalamus	(Intercept)	0.237	0.656	0.36	$7.18 \cdot 10^{-1}$	
	scan.age	0.036	0.013	2.85	$4.69 \cdot 10^{-3}$	**
	low	0.479	0.391	1.23	$2.21 \cdot 10^{-1}$	
	med	0.362	0.354	1.02	$3.08 \cdot 10^{-1}$	
	high	0.536	0.491	1.09	$2.76 \cdot 10^{-1}$	
	gender.male	−0.206	0.328	−0.63	$5.31 \cdot 10^{-1}$	

No significant effect of disease progression, represented with CAG-Age Product (CAP) group, on automated method's reliability is detected at the level 0.001: $CV = \beta_0 + \beta_1 AGE + \beta_2 CAP + \beta_3 Gender$. Significant level of p-value is marked with [***], [**], [*], and [.] for < 0.001, < 0.01, < 0.05, and < 0.1, respectively.

6.6 Real World Data Application

We also evaluated *BRAINSCut* ability to process a wide range of data by applying one of large-scale multicenter data, MR data collected from PREDICT-HD study [127], to drive six subcortical structures. The software robustness and segmentation quality were quantified with a success ratio through out the visual inspection. The completion ratio of the software, the proportion of scans that completed without error, was small on PREDICT-HD [127] data. The quality of derived subcortical structures is visually rated according to the provided guideline [3] and results, in terms of these three level grading, are shown in Table 6.7. The segmentation quality rated at poor level was substantially small in number ($< 6\%$) across over 3000 scan sessions. Samples of each rated as poor, reasonable, and good quality of segmentation results are also shown in Figure 6.8 as well as Appendix FigureA.8, A.9, and A.10 with more details.

BRAINSCut result was competitive with the most commonly used software, FreeSurfer. Examples of *BRAINSCut* and *FreeSurfer* segmentation of six subcortical structures 6.9 indicated the similarity between the results between the two methods.

[3] Derived six subcortical structures from *BRAINSCut* are rated by three independent experts for their qualities. The rating is based on three levels: 0=poor, 1=reasonable, and 2=good. Poor, reasonable, good quality segmentation indicates usable with edits required, sufficient quality but would benefit from small edit, and perfect as is, respectively, for further quantitative analysis.

Table 6.7: Ration of Visual investigation scores of six subcortical structures from two large-scale studies, Predict-HD [127] (Upper) and TrackOn [161] (Bottom), are summarized.

PREDICT HD	accumben		caudate		putamen		globus		thalamus		hippocampus	
score ($n = 2228$)	left	right	left	right	left	right	left	right	left	right	left	right
0(poor)	*0.007	*0.005	*0.044	*0.058	*0.057	*0.078	*0.041	*0.052	*0.018	*0.027	*0.016	*0.011
1	0.065	0.088	0.177	0.255	0.161	0.201	0.096	0.123	0.074	0.105	0.089	0.079
2(good)	0.928	0.907	0.779	0.686	0.783	0.721	0.864	0.825	0.908	0.868	0.895	0.911

Track On	accumben		caudate		putamen		globus		thalamus		hippocampus	
score ($n = 782$)	left	right	left	right	left	right	left	right	left	right	left	right
0(poor)	*0.028	*0.040	*0.030	*0.052	*0.058	*0.091	*0.095	*0.085	*0.009	*0.015	*0.037	*0.033
1	0.103	0.119	0.077	0.122	0.136	0.143	0.183	0.187	0.005	0.022	0.078	0.081
2(good)	0.869	0.841	0.893	0.827	0.806	0.766	0.722	0.729	0.985	0.963	0.885	0.886

Three experts rated segmentation quality on three level basis: 0=*poor*, 1=*reasonable with minor manual intervention*, and 3= *good* quality of automated segmentations. Our method is applied on over 3000 scans and failure rate is less then 10% for all the structures. The table also reveal that the higher failure rate for globus than other structures. Samples of each poor, reasonable, and good segmentation from BRAINSCut are given in Figure 6.8 and also in Appendix Figure A.8, A.9, and A.10, including more details.

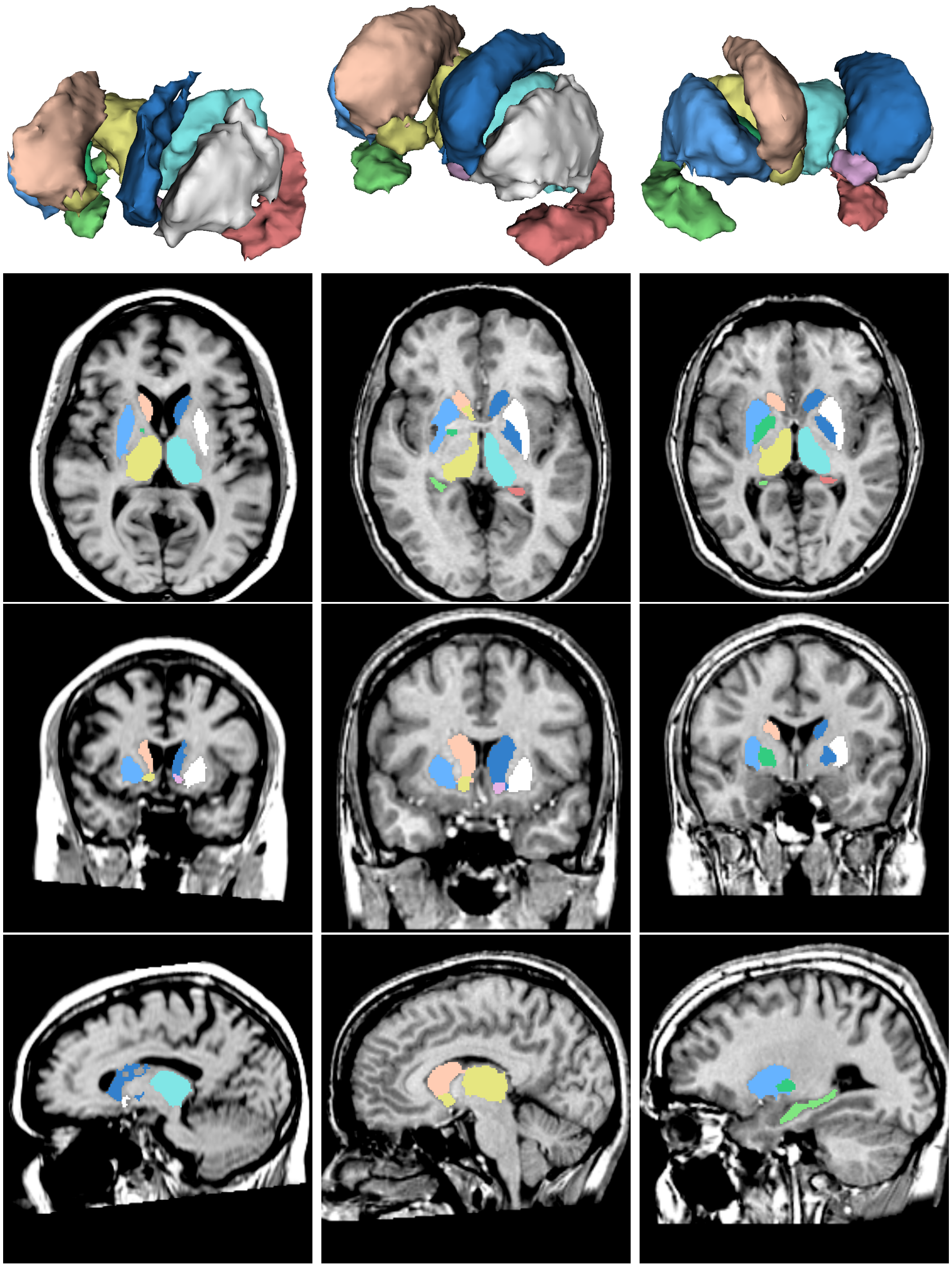


Figure 6.8: Subcortical segmentation examples from *BRAINSCut.* From top to bottom, images of each row corresponds to be rated as ‘poor’, ‘reasonable’, and ‘good’ via the visual inspection.

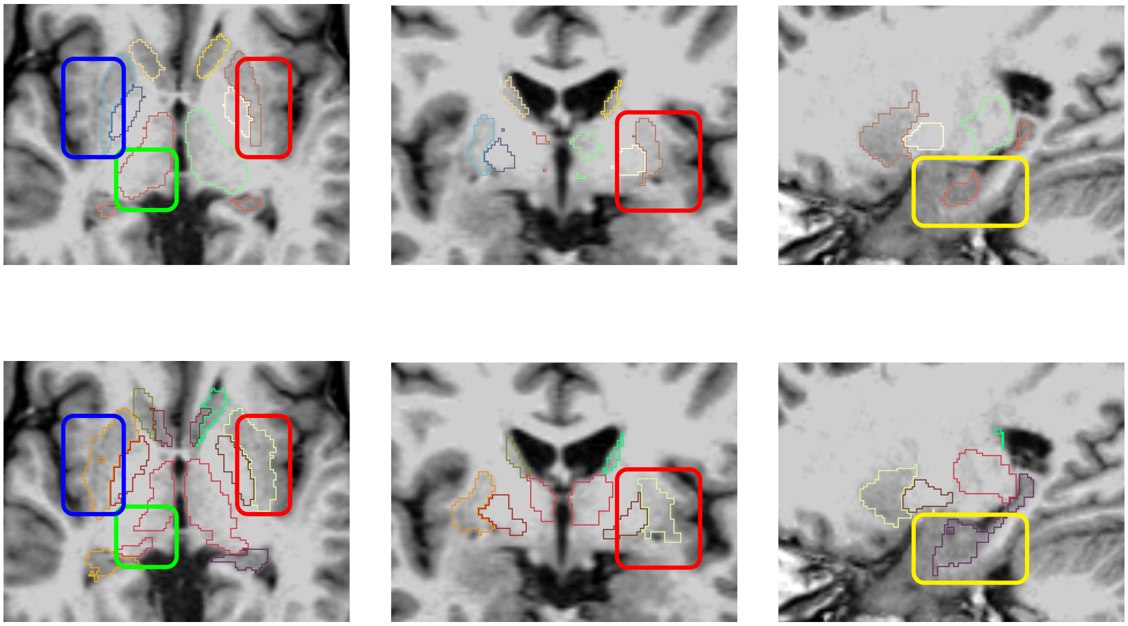


Figure 6.9: Contrasted segmentation results between *FreeSurfer* (bottom row) and the proposed method, *BRAINSCut* (upper row), to assess validity of segmentation results in terms of MRI-driven shape. From the left to right column, segmentations are presented in axial, coronal, and sagittal view. Obvious differences between two methods are marked with red, blue, green, and yellow boxes for right and left putamen, thalamus, and hippocampus, respectively. It is particularly noticeable that the delineation of putamen (red and blue boxes) is well over the actual physiological definition and inclusive to the claustrum (See the Appendix Figure A.1 for the claustrum) The FreeSurfer often segments volumes of subcortical structures over-inclusively, that may degrades the quality of consequence analysis, such as longitudinal shape changes or precise volumetric change access along the disease progression.

6.7 Summary and Conclusion

Quality of the segmentation results were evaluated in two steps. First validation was provided with the high correspondence measure to gold standards from manual segmentations, including relative overlap (RO), Hausdorff distance (HD), ICCs, and other measures that we provided from the cross-validation experiments. 10-fold cross-validation provides correspondences from hold-out, and so unseen data, validity of segmentation results were tested. The second part of validity assessments was the visual inspection phase. All the processed segmentation results are visually inspected by human experts and rated as excellent/good/bad. The visual inspection results showed that less than 10% failure ratio over the 3000 scans. Good reliability of the segmentation framework for six subcortical structures was evaluated two data sets: 1) THP and 2) repeated scans of same subject within a week. While THP data set provided degrees of intra- and inter-center reliability, repeated scans demonstrated stability of the measurement between two time points, where negligible biological change is expected. In addition, we also tested effect of disease progression on the measurement stability by using same data set. Controlling for age and gender, we investigated if there exists any group effects based on CAP score, a degree of disease progression of HD, on the scan-rescan reliability.

CHAPTER 7
CONCLUSION AND FUTURE WORK

This paper describes a segmentation framework to delineate the brain subcortical structures consistently from large-scale multicenter MR data set. Excellent robustness and validity of BRAINSCut are achieved by employing multiple safety devices that are described through out five chapters:

1. An improved bias-correction algorithm through out iterative process between bias-correction, registration, and tissue classification (Chapter 2),
2. An proper choice of machine algorithm (Chapter 3), a region specific normalization (Chapter 4), and a custom set of feature-enhanced images (Chapter 5), and
3. A series of validation study that occurs repeatedly together with the software development to make sure the robustness and reliability (Chapter 6).

We underlined promising results that were obtained from the previous investigation on the segmentation framework in Chapter 1, which is based on the machine-learning techniques. The segmentation framework constructed upon ANN resulted in very high segmentation correspondence to manual traces by using multi-modal MRIs. The segmentation framework trained on only 24 data was successfully applied on hundreds of data that covers a wide variety of brain morphologies from healthy normal control to pre-HD patients. The plausibility of results leads us to investigate machine-learning based segmentation framework for large-scale multicenter MR data.

Chapter 2 empirically showed the enhanced bias-correction algorithm improved the segmentation accuracy. The segmentation accuracy was obviously improved when it measured in the correspondence to the manual traces in terms of ICC. This bias-corrected MRI image is the product of the iterative optimization framework between bias-correction, registration, and tissue classification, that has been enhanced its tissue classification performance and robustness against large-scale data set. The improvement of iterative bias-correction algorithm was shown via two data set: 1) the improved tissue classification accuracy from the simulated MRI data, BrainWeb, experiment with a provided ground truth, and 2) a higher success rate with visual inspection by human experts on the in-vivo application. The robust pre-processing, bias-correction step, benefits the automated segmentation framework.

A random forest algorithm achieved the best segmentation accuracy among 12 variations of machine-learning algorithm as described in Chapter 3. The experimental result suggested that the superiority of the random forest algorithm in terms of accuracy and generalizability. Through out the subject-basis 10-cross validation study, the best segmentation accuracy achieved with the random forest algorithm for all six subcortical structures. The experiment also confirms the generalizability of the random forest while ANNs over-fitting issue was observed. The random forest algorithm was integrated into the segmentation framework for large-scale data processing.

An investigation of the region-specific normalization occurred in Chapter 4. This normalization approach has upgraded the subcortical segmentation accuracy with statistical significance. 11 variations of intensity normalization functions were plugged into the regions-specific normalization framework and result showed the benefits of each normalization functions. Two normalization functions are selected for the final segmentation framework based on their statistical significance: a linear (min/max) and an IQR-based transformation function. IQR-based function was preferred by all the structures but caudate nucleus. The relative spatial location of caudate nucleus, which is the only structure that adjacent to CSF, favored the linear (min/max) normalization function.

A feature image, a summed image of gradient magnitude of T1- and T2-weighted images ($\mathcal{I}_{SG}$), showed the best improvement for the subcortical segmentation in Chapter 5. Two groups of feature-enhanced images, total of eight feature-enhanced images are hierarchically investigated and the series of experiments favored $\mathcal{I}_{SG}$. Feature forward selection method was adapted in the hierarchical investigation of eight feature-enhanced images and each comparative experiment performed a 10-fold cross validation. The superiority of $\mathcal{I}_{SG}$ was confirmed through out a set of experiments and finalized our custom feature set $\{\mathcal{I}_{SG}, \mathcal{I}_{T1}, \mathcal{I}_{T2}\}$ (or $\{\mathcal{I}_{GM(T1)}, \mathcal{I}_{T1}\}$ for uni-modal scan).

Finally, in Chapter 6 the segmentation framework displayed an excellent reliability against a wide range of disease status, a high success ratio, and a great validity form the visual inspection. The result was visually compatible to the state of art in the field, FreeSurfer. The validation is done via two sets of in-vivo MRI data: 1) THP for the multicenter reliability, and 2) the MRI data from the PREDICT-HD study for repeated measure reliability. Both reliability studies presented no statistically significant evidence of measurement differences either between centers or disease status.

The lack of robust automated segmentation methods results in tedious manual labor. The reduction in operator time (6-10 hours to 5-10 minutes) makes it practical to consider the integration of computerized segmentation into a large-scale clinical data analysis. The result of this paper suggests that the automated segmentation framework, based on machine-learning techniques, operates robustly on the large-scale multicenter MR data. This dissertation described a collaborative effort that incorporating multiple medical engineering techniques. The robustness, which is essential in the design of efficient clinical trials, is accomplished via these carefully engineered safety devices.

Limited comparative studies to other available tools in the field were performed. A comprehensive comparative study among available segmentation tools in the field is required to ensure which approach is the best suite for a specific research setting. Label-fusion/propagation-based segmentation method is one of emerging approach in recent years [140, 75, 96, 170, 169, 179, 29, 68, 57, 111]. A extensive study between those available and promising techniques will guide us in new segmentation improvements. Future studies can examine the possibility that Machine-learning based segmentation on whole brain segmentation ([81, 140]). Whole brain segmentation can benefits the accuracy by explicitly penalizing the possibility of mismatch between structures of interest and background tissues.

APPENDIX

A.1 Notation

Table A.1: Notation used in this report

Data notation	
Image	$\mathcal{I}$
Voxel Location	$i \in \{1, \cdots, n\}$
Feature Vector	$f_i \in \mathbf{F}$
Output Vector	$y_i \in \mathbf{N}$
Sample data	$\mathcal{S}_{n_s}$ for finit number n_s
Population Data	$\mathcal{X}$
Model	$\mathbb{M}$
True prediction error	e
Apparent error	e_a

A.2 Backpropagation Algorithm Derivation

For the target output t_i and the node output o_i for each node i, ANN's backpropagation algorithm tries to minimize error for the entire net:

$$E = \sum_i E_i,$$

where E_i is the error at each node at i:

$$E_i = \frac{1}{2}(t_i - o_i)^2.$$

The node output o_i depends on its connected input node and weight:

$$o_i = sigmoid(net_i) = sigmoid(\sum_j w_{ij} o_i),$$

where w_{ij} is a weight from node i to j. Then, ANN's learning is to reduce E by adjusting weights of w_{ij}:

$$\begin{aligned}
\Delta w_{ij} &\propto -\frac{\partial E}{\partial w_{ij}} && (1)\\
&\propto -\frac{\partial E}{o_i} \cdot \frac{o_i}{\partial w_{ij}} && (2)\\
&\propto -\frac{\partial}{\partial o_i}\left(\frac{1}{2}\sum_k (t_k - o_k)^2\right) \cdot \frac{\partial}{\partial w_i j} sigmoid(net_i) && (3)\\
&\propto (t_i - o_i) o_i (1 - o_i) && (4)\\
&= \eta \delta_i o_i. && (5)
\end{aligned}$$

η is a learning rate, which is a fixed constant for the entire net.

A.3 Random Forest: Generalizability

The proof is from [20]. Let X be a data space, Y a set of classes, and (X,Y) the space of correct pairings of data points with class. Let Θ be the distribution that directs the randomization of individual tree classifiers h_θ in the forest. Define a function $\hat{j}(\{x, y\})$ which outputs the class receiving the highest proportion of votes, other than the correct class y, when x is classified by the random forest. That is,

$$\hat{j}(\{x, y\}) = argmax_{j \neq y} P_\Theta(h_\theta(x) = j).$$

Then the *margin function* of the random forest for a point x,y is

$$mr(\{x, y\}) = P_\Theta(h_\theta(x) = y) - P_\Theta(h_\theta(x) = \hat{j}(\{x, y\})),$$

the extent to which the forest prefers (or fail to prefer) the correct classification for x over the closest competing classification. Since the margin function of a point the forest gets right will be negative, the generalization error PE^* is

$$PE^* = P_{(X,Y)}(mr(\{x, y\} < 0).$$

A reasonable measure of the strength s of the forest is

$$s = E_{(X,Y)}[mr(\{x, y\})].$$

Assuming that s is non-negative (a safe assumption - if s is negative the random forest could be beaten by a coin flip and is not worth studying), Chebyshev's inequality allows us to bound the generalization error:

$$\begin{aligned} PE^* &= P_{(X,Y)}(mr(\{x, y\}) < 0) \\ &\leq P_{(X,Y)}(|mr(\{x, y\}) - s| \geq s) \\ &\leq \frac{var_{(X,Y)}(mr(\{x, y\})}{s^2} \end{aligned}$$

A.4 Additional Tables and Figures

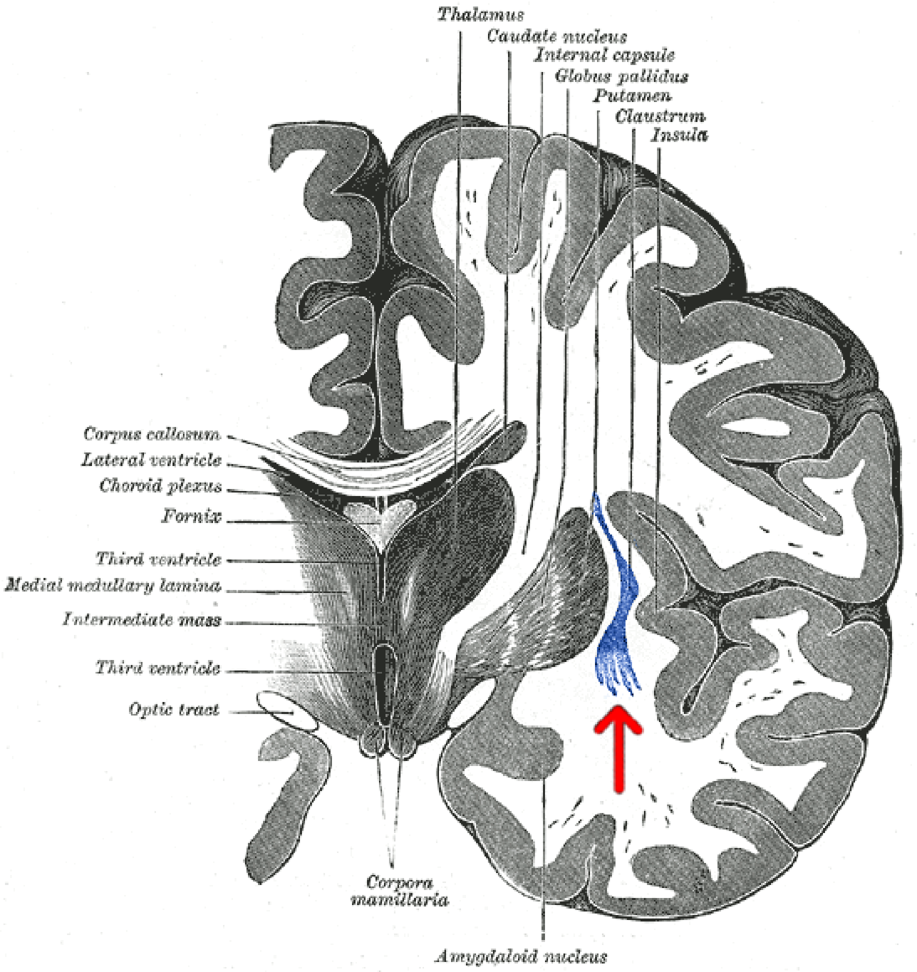


Figure A.1: Claustrum, an adjacent structure to putamen, that is often misclassifed to putamen from an automated segmentation method. The picture is from `http://en.wikipedia.org/wiki/File:Gray_718-emphasizing-claustrum.png`

Table A.2: Screening study results for *hippocampus* contrasting performance of 12-machine learning approaches on the identical MR segmentation data in WEKA.

Location	Sensitivity	Specificity	Precision	F-Measure	AUC	Sensitivity	Specificity	Precision	F-Measure	AUC
			Majority Classifier					**Bagging**		
Left	0	0	0	0	0	0.783	0.032	0.81	0.796	* **0.977**
Right	0	0	0	0	0	0.762	0.031	0.801	0.781	* **0.975**
None	1	1	0.712	1	0.832	* **0.925**	0.227	* **0.91**	* **0.917**	* **0.943**
Avg.	0.712	0.712	0.507	0.712	0.592	0.881	0.171	0.88	0.88	* **0.953**
			Naïve bayes					**kNN** ($k = 1$)		
Left	0.496	0.142	0.379	0.496	0.429	0.835	0.041	0.781	0.835	0.807
Right	0.442	0.154	0.317	0.442	0.369	0.811	0.041	0.761	0.811	0.785
None	0.783	0.186	* **0.912**	0.783	0.843	* **0.901**	0.177	* **0.927**	* **0.901**	* **0.914**
Avg.	0.693	0.175	0.75	0.693	0.715	0.879	0.138	0.882	0.879	0.88
			SVM					**kNN** ($k = 10$)		
Left	0	0	0	0	0	0.835	0.041	0.781	0.835	0.807
Right	0	0	0	0	0	0.811	0.041	0.761	0.811	0.785
None	1	1	0.712	1	0.832	* **0.901**	0.177	* **0.927**	* **0.901**	* **0.914**
Avg.	0.712	0.712	0.507	0.712	0.592	0.879	0.138	0.882	0.879	0.88
			AdaBoost					**kNN** ($k = 20$)		
Left	0	0	0	0	0	0.88	0.037	0.805	0.88	0.841
Right	0	0	0	0	0	0.848	0.034	0.803	0.848	0.825
None	1	1	0.712	1	0.832	* **0.915**	0.135	* **0.944**	* **0.915**	* **0.929**
Avg.	0.712	0.712	0.507	0.712	0.592	* **0.9**	0.107	* **0.903**	* **0.9**	* **0.901**
			ANN ($HN = 20$)					**Random forest** ($NT = 10$)		
Left	0.854	0.029	0.836	0.854	0.845	0.747	0.025	0.837	0.747	0.79
Right	0.806	0.031	0.806	0.806	0.806	0.723	0.026	0.82	0.723	0.768
None	* **0.927**	0.169	* **0.931**	* **0.927**	* **0.929**	* **0.939**	0.264	0.898	* **0.939**	* **0.918**
Avg.	0.899	0.129	* **0.9**	0.899	0.899	0.88	0.196	0.878	0.88	0.878
			ANN ($HN = 60$)					**Random forest** ($NT = 25$)		
Left	0.853	0.027	0.847	0.853	0.85	0.794	0.028	0.83	0.794	0.812
Right	0.839	0.03	0.817	0.839	0.828	0.77	0.028	0.814	0.77	0.791
None	* **0.931**	0.154	* **0.937**	* **0.931**	* **0.934**	* **0.932**	0.218	* **0.914**	* **0.932**	* **0.923**
Avg.	* **0.907**	0.118	* **0.907**	* **0.907**	* **0.907**	0.889	0.163	0.887	0.889	0.888

The superiocity of Random Forest algorithm and ANN are well supported by the series of experiemnts reported in Table 3.2, A.3, A.4, and this one. Five measurements are reported for left (L), right (R), background (Bg) of caudate, and average of all three regions (Avg): *sensitivity*, *specificity*, *precision*, *F-measure*, and *area under the curve* (AUC). A metric with preferable performance (> 0.9) are written in bold with * mark: Bagging, variation of ANN ($HN = 20$ and 60), kNN ($k = 1$, 10, and 20), and random forest ($NT = 10$ and 25). The full related description is in Chapter 3.4.

Table A.3: Screening study results for *putamen* contrasting performance of 12-machine learning approaches on the identical MR segmentation data in WEKA.

Location	Sensitivity	Specificity	Precision	F-Measure	AUC	Sensitivity	Specificity	Precision	F-Measure	AUC
			Majority Classifier					**Bagging**		
Left	0	0	0	0	0.5	0.776	0.035	0.818	0.797	*** 0.974**
Right	0	0	0	0	0.5	0.782	0.036	0.817	0.799	*** 0.974**
None	1	1	0.662	0.796	0.5	*** 0.911**	0.221	0.89	*** 0.9**	*** 0.935**
Avg.	0.662	0.662	0.438	0.527	0.5	0.866	0.158	0.865	0.866	*** 0.949**
			Naïve bayes					**kNN** ($k = 1$)		
Left	0.079	0.02	0.437	0.133	0.733	0.8	0.048	0.771	0.785	0.876
Right	0.42	0.176	0.33	0.37	0.714	0.804	0.051	0.766	0.785	0.878
None	0.853	0.556	0.75	0.798	0.76	0.876	0.198	0.897	0.886	0.839
Avg.	0.649	0.401	0.626	0.613	0.747	0.851	0.147	0.853	0.852	0.852
			SVM					**kNN** ($k = 10$)		
Right	0.407	0.037	0.693	0.513	0.685	0.788	0.031	0.836	0.811	*** 0.976**
None	*** 0.909**	0.588	0.751	0.823	0.661	0.865	0.05	0.781	0.821	*** 0.976**
Left	0.417	0.035	0.705	0.524	0.691	0.898	0.173	*** 0.91**	*** 0.904**	*** 0.944**
Avg.	0.741	0.401	0.734	0.72	0.67	0.874	0.128	0.876	0.874	*** 0.955**
			AdaBoost					**kNN** ($k = 20$)		
Left	0	0	0	0	0.199	0.799	0.035	0.82	0.809	*** 0.978**
Right	0	0	0	0	0.801	0.854	0.047	0.791	0.821	*** 0.978**
None	1	1	0.662	0.796	0.503	0.897	0.173	*** 0.91**	*** 0.904**	*** 0.945**
Avg.	0.662	0.662	0.438	0.527	0.503	0.873	0.129	0.875	0.874	*** 0.956**
			ANN ($HN = 20$)					**Random forest** ($NT = 10$)		
Left	0.847	0.051	0.768	0.806	*** 0.972**	0.742	0.028	0.842	0.789	*** 0.968**
Right	0.82	0.048	0.777	0.798	*** 0.97**	0.822	0.045	0.791	0.807	*** 0.969**
None	0.874	0.166	*** 0.911**	0.892	*** 0.929**	*** 0.909**	0.217	0.891	*** 0.9**	*** 0.931**
Avg.	0.861	0.127	0.864	0.862	*** 0.943**	0.866	0.156	0.866	0.865	*** 0.943**
			ANN ($HN = 60$)					**Random forest** ($NT = 25$)		
Left	0.859	0.042	0.805	0.831	*** 0.978**	0.792	0.032	0.832	0.812	*** 0.976**
Right	0.846	0.043	0.803	0.824	*** 0.976**	0.802	0.033	0.832	0.817	*** 0.976**
None	0.894	0.148	*** 0.922**	*** 0.908**	*** 0.944**	*** 0.918**	0.203	0.898	*** 0.908**	*** 0.943**
Avg.	0.88	0.112	0.882	0.881	*** 0.955**	0.877	0.145	0.876	0.876	*** 0.954**

The superiocity of Random Forest algorithm and ANN are well supported by the series of experiemnts reported in Table 3.2, A.2, A.4, and this one. Five measurements are reported for left (L), right (R), background (Bg) of caudate, and average of all three regions (Avg): *sensitivity*, *specificity*, *precision*, *F-measure*, and *area under the curve* (AUC). A metric with preferable performance (> 0.9) are written in bold with * mark: Bagging, variation of ANN ($HN = 20$ and 60), kNN ($k = 1$, 10, and 20), and random forest ($NT = 10$ and 25). The full related description is in Chapter 3.4.

Table A.4: Screening study results for *thalamus* contrasting performance of 12-machine learning approaches on the identical MR segmentation data in WEKA

Location	Sensitivity	Specificity	Precision	F-Measure	AUC	Sensitivity	Specificity	Precision	F-Measure	AUC
			Majority Classifier					**Bagging**		
Left	0	0	0	0	0.5	0.868	0.037	0.863	0.865	* **0.983**
Right	0	0	0	0	0.5	0.857	0.033	0.857	0.857	* **0.983**
None	1	1	0.6	0.75	0.5	* **0.907**	0.137	* **0.909**	* **0.908**	* **0.956**
Avg.	0.6	0.6	0.36	0.45	0.5	0.889	0.096	0.889	0.889	* **0.967**
			Naïve bayes					**kNN** $(k = 1)$		
Left	0.745	0.427	0.319	0.447	0.731	0.863	0.043	0.843	0.853	* **0.91**
Right	0.221	0.145	0.261	0.239	0.671	0.86	0.038	0.838	0.849	* **0.911**
None	0.504	0.11	0.873	0.639	0.795	0.891	0.138	* **0.907**	0.899	0.876
Avg.	0.502	0.184	0.641	0.523	0.758	0.88	0.099	0.88	0.88	0.89
			SVM					**kNN** $(k = 10)$		
Left	0.726	0.078	0.715	0.72	0.824	0.87	0.03	0.886	0.878	* **0.984**
Right	0.642	0.064	0.699	0.67	0.789	0.858	0.028	0.876	0.867	* **0.985**
None	0.811	0.313	0.796	0.803	0.749	* **0.923**	0.136	* **0.911**	* **0.917**	* **0.963**
Avg.	0.762	0.217	0.76	0.761	0.772	0.899	0.093	0.899	0.899	* **0.972**
			AdaBoost					**kNN** $(k = 20)$		
Left	0	0	0	0	0.813	0.882	0.034	0.873	0.877	* **0.986**
Right	0	0	0	0	0.19	0.867	0.031	0.865	0.866	* **0.986**
None	1	1	0.6	0.75	0.52	* **0.912**	0.125	* **0.916**	* **0.914**	* **0.964**
Avg.	0.6	0.6	0.36	0.45	0.52	0.897	0.088	0.897	0.897	* **0.973**
			ANN ($HN = 20$)					**Random forest** ($NT = 10$)		
Left	0.817	0.039	0.85	0.833	* **0.973**	0.828	0.029	0.883	0.855	* **0.978**
Right	0.839	0.039	0.832	0.836	* **0.976**	0.821	0.027	0.877	0.848	* **0.978**
None	0.896	0.172	0.887	0.891	* **0.935**	* **0.925**	0.175	0.888	* **0.906**	* **0.951**
Avg.	0.869	0.119	0.869	0.869	* **0.951**	0.885	0.116	0.885	0.884	* **0.962**
			ANN ($HN = 60$)					**Random forest** ($NT = 25$)		
Left	0.836	0.03	0.881	0.858	* **0.98**	0.867	0.034	0.873	0.87	* **0.984**
Right	0.865	0.031	0.865	0.865	* **0.983**	0.858	0.03	0.868	0.863	* **0.984**
None	* **0.918**	0.15	* **0.902**	* **0.91**	* **0.952**	* **0.915**	0.137	* **0.909**	* **0.912**	* **0.96**
Avg.	0.89	0.103	0.89	0.89	* **0.964**	0.894	0.095	0.894	0.894	* **0.97**

The superiocity of Random Forest algorithm and ANN are well supported by the series of experiemnts reported in Table 3.2, A.3, A.2, and this one. Five measurements are reported for left (L), right (R), background (Bg) of caudate, and average of all three regions (Avg): *sensitivity*, *specificity*, *precision*, *F-measure*, and *area under the curve* (AUC). A metric with preferable performance (> 0.9) are written in bold with * mark: Bagging, variation of ANN (HN = 20 and 60), kNN (k = 1, 10, and 20), and random forest ($NT = 10$ and 25). The full related description is in Chapter 3.4.

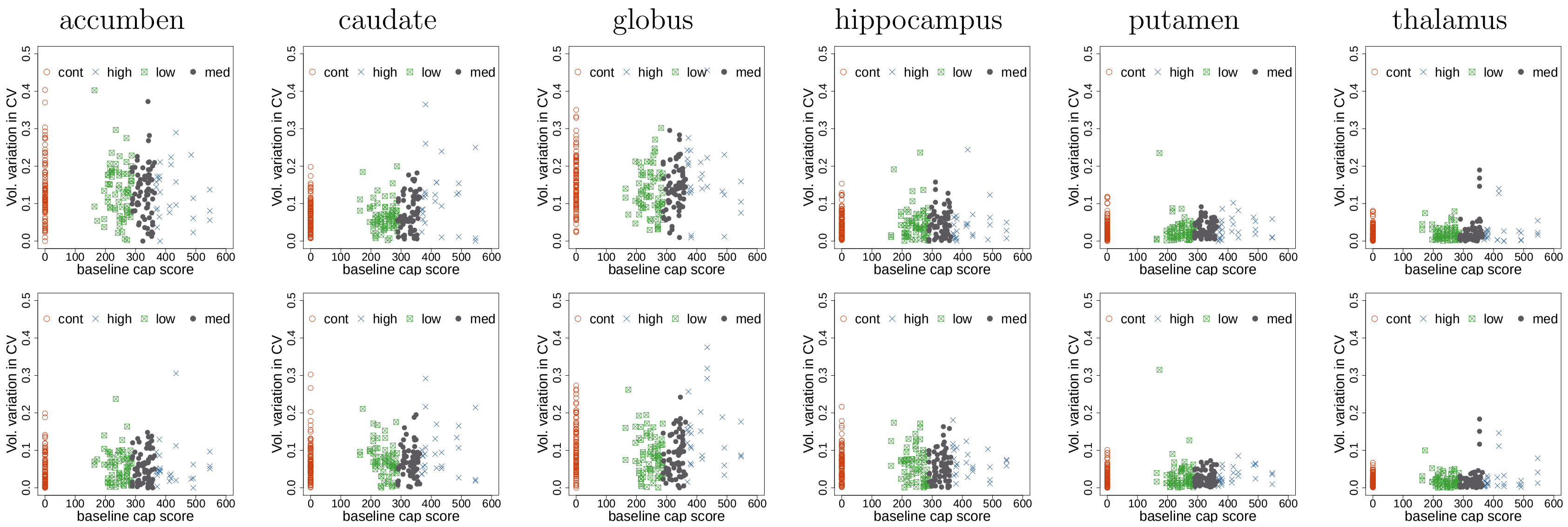


Figure A.2: Scatter plots of CV across CAP group for each ROI showing no specific pattern of measurement variation, encoded in CV, neither by CAP group nor ROI. Left (upper) and right (bottom) structures are plotted separately.

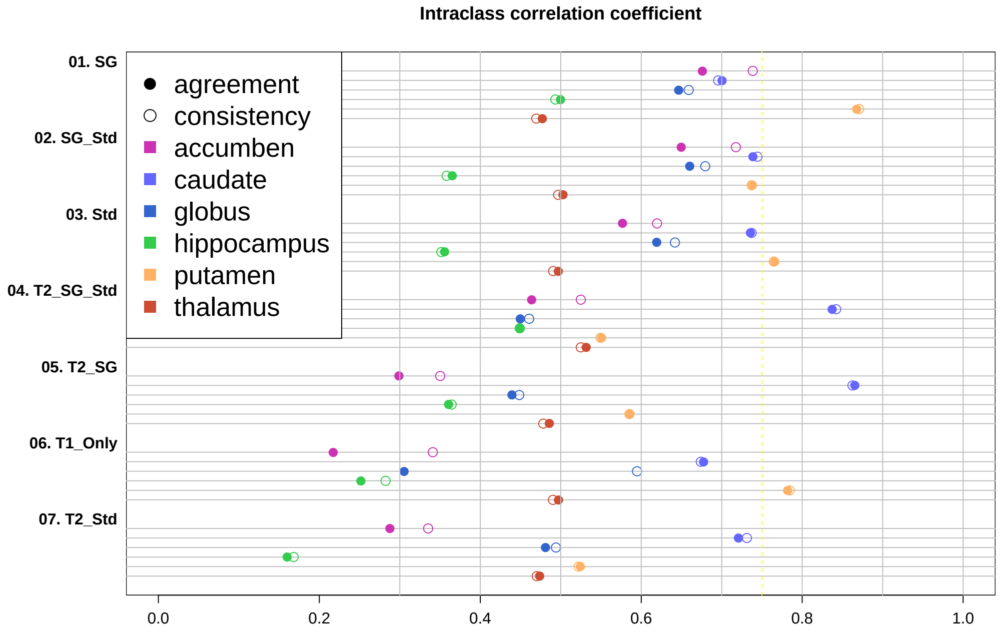


Figure A.3: Comparing Seven Combinations of Candidate Feature Set without No Normalization Applied $ICC(A)$ (solid circle) and $ICC(C)$ (empty circle) dot graph is shown for seven combinations of the most promising feature-enhanced images without normalization (None). All six structures are tested and plotted with different colors. ICC's lower bound suggested by Shrout [145] also presented as a red line. Seven combinations marked on the left-hand side are ranked by its average performance over six structures from the top. That is, *'01 SG'* and *'02. SG_Std'* presented top three best average performance over six structures based on ICCs. Also note that this particular plot, which used no intensity normalization, is more spread out than other experiments that employed a normalization. **In other words, without normalization, the segmentation performance is more sensitvie to the choice of the feature set.** Also note that the T2-weighted image together with standard deviation (std) lowered the performance compared to the one that used T1-weighted image only (T1_Only)

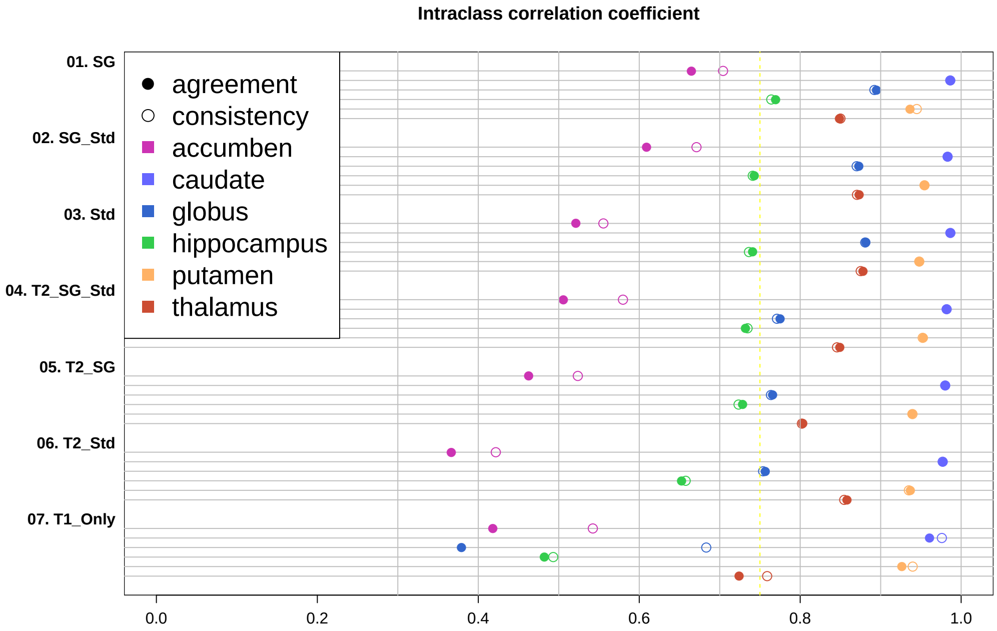


Figure A.4: Comparing Seven Combinations of Candidate Feature Set Linear (min/max) Normalization Applied . $ICC(A)$ (solid circle) and $ICC(C)$ (empty circle) dot graph is shown for seven combinations of the most promising feature-enhanced images without normalization (None). All six structures are tested and plotted with different colors. ICC's lower bound suggested by Shrout [145] also presented as a red line. Seven combinations marked on the left-hand side are ranked by its average performance over six structures from the top. That is, *'01 SG'* and *'02. SG_Std'* presented top three best average performance over six structures based on ICCs.

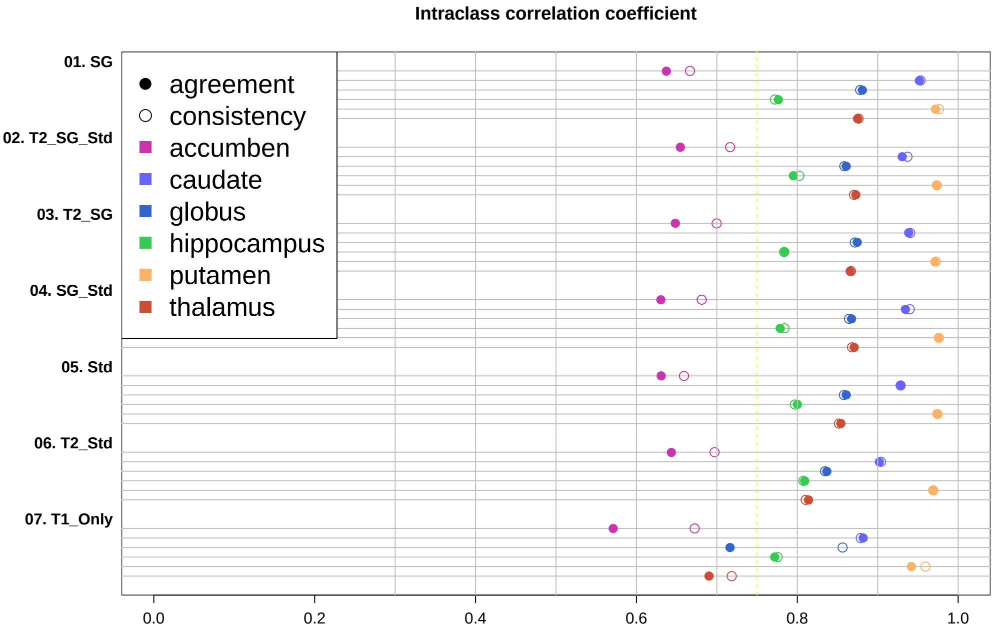


Figure A.5: Comparing Seven Combinations of Candidate Feature Set IQR-based Normalization Applied . $ICC(A)$ (solid circle) and $ICC(C)$ (empty circle) dot graph is shown for seven combinations of the most promising feature-enhanced images without normalization (None). All six structures are tested and plotted with different colors. ICC's lower bound suggested by Shrout [145] also presented as a red line. Seven combinations marked on the left-hand side are ranked by its average performance over six structures from the top. That is, *'01 SG'* and *'02. T2_SG_Std'* presented top three best average performance over six structures based on ICCs.

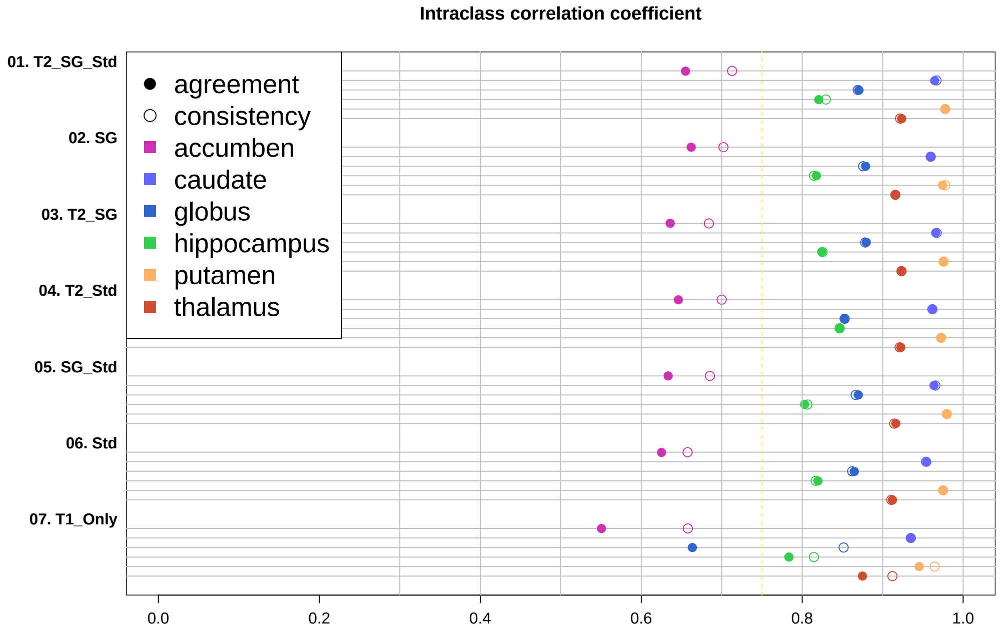


Figure A.6: Comparing Seven Combinations of Candidate Feature Set IQR-based Normalization Applied . $ICC(A)$ (solid circle) and $ICC(C)$ (empty circle) dot graph is shown for seven combinations of the most promising feature-enhanced images without normalization (None). All six structures are tested and plotted with different colors. ICC's lower bound suggested by Shrout [145] also presented as a red line. Seven combinations marked on the left-hand side are ranked by its average performance over six structures from the top. That is, *'01 T2 SG Std'* and *'02. SG'* presented top three best average performance over six structures based on ICCs.

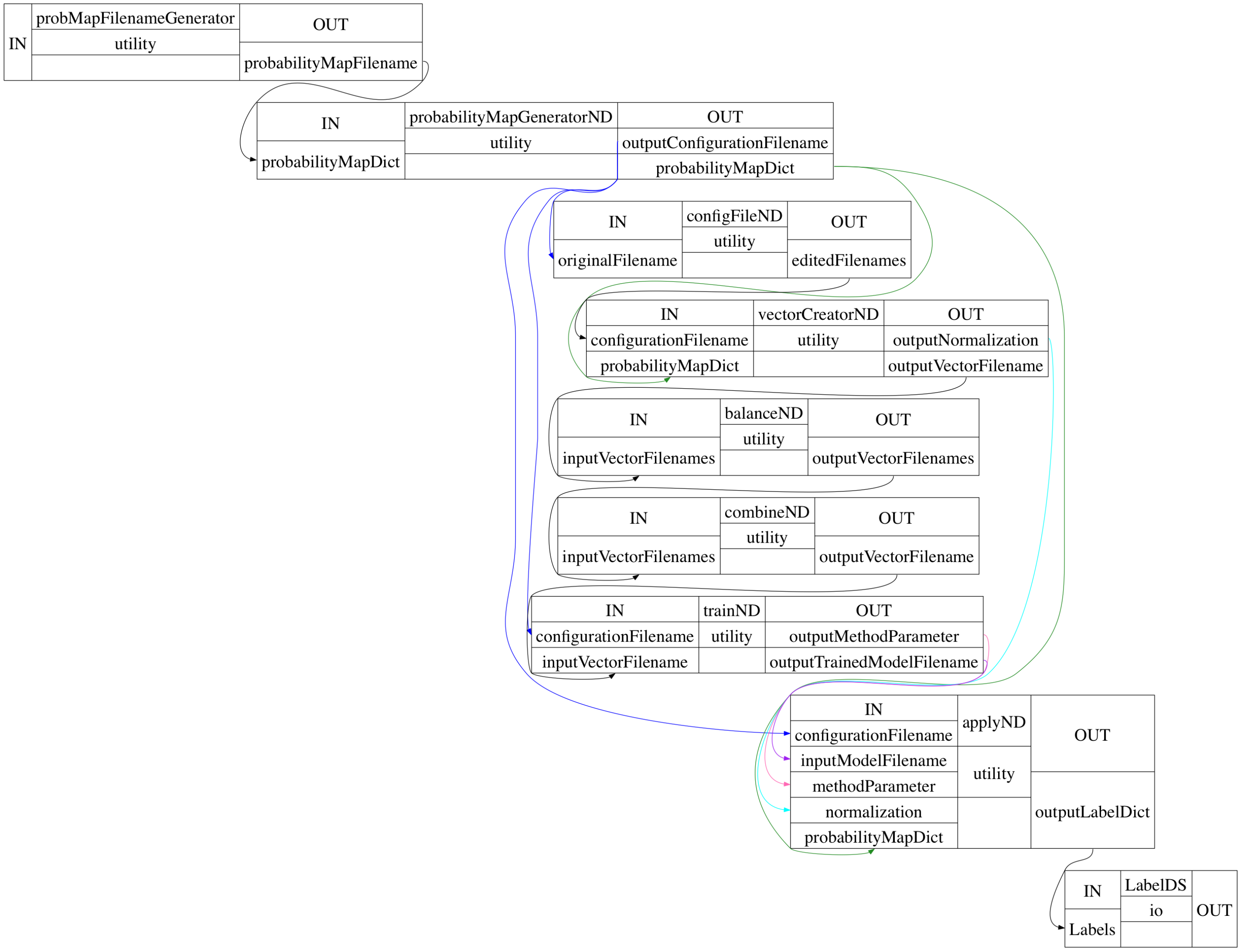
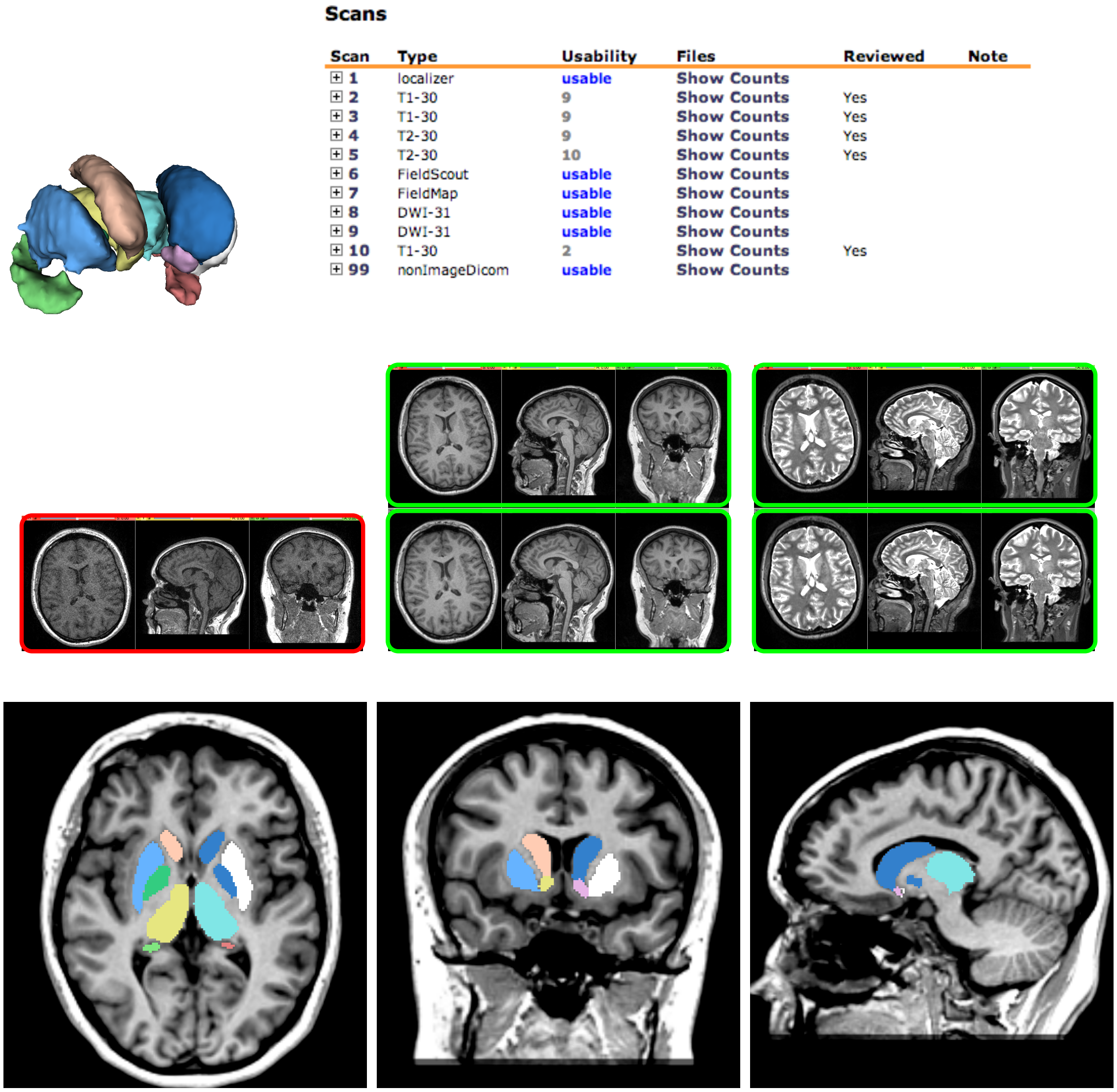


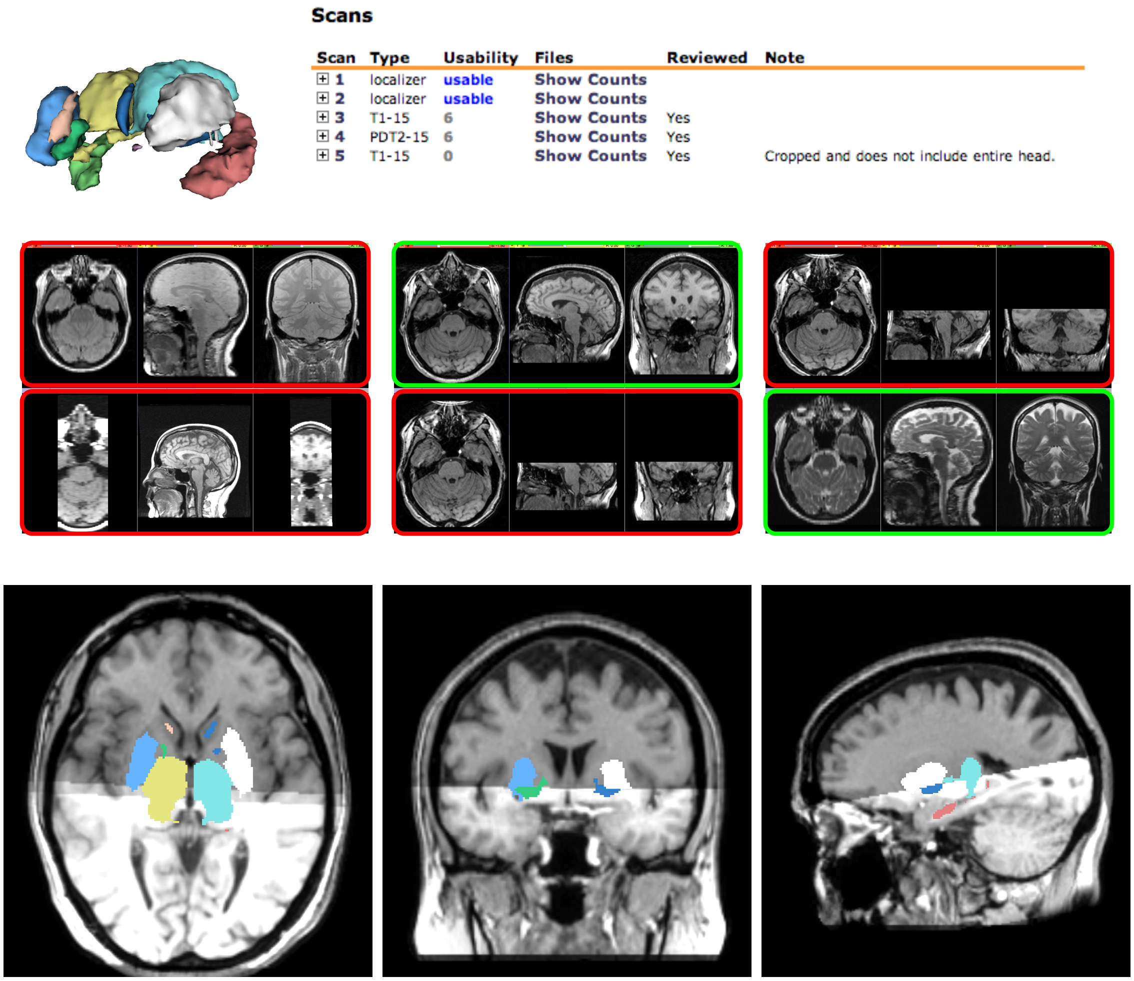
Figure A.7: Cross-validation NiPype framework in details with their I/O connections



Scans

Scan	Type	Usability	Files	Reviewed	Note
⊞ 1	localizer	usable	Show Counts		
⊞ 2	T1-30	9	Show Counts	Yes	
⊞ 3	T1-30	9	Show Counts	Yes	
⊞ 4	T2-30	9	Show Counts	Yes	
⊞ 5	T2-30	10	Show Counts	Yes	
⊞ 6	FieldScout	usable	Show Counts		
⊞ 7	FieldMap	usable	Show Counts		
⊞ 8	DWI-31	usable	Show Counts		
⊞ 9	DWI-31	usable	Show Counts		
⊞ 10	T1-30	2	Show Counts	Yes	
⊞ 99	nonImageDicom	usable	Show Counts		

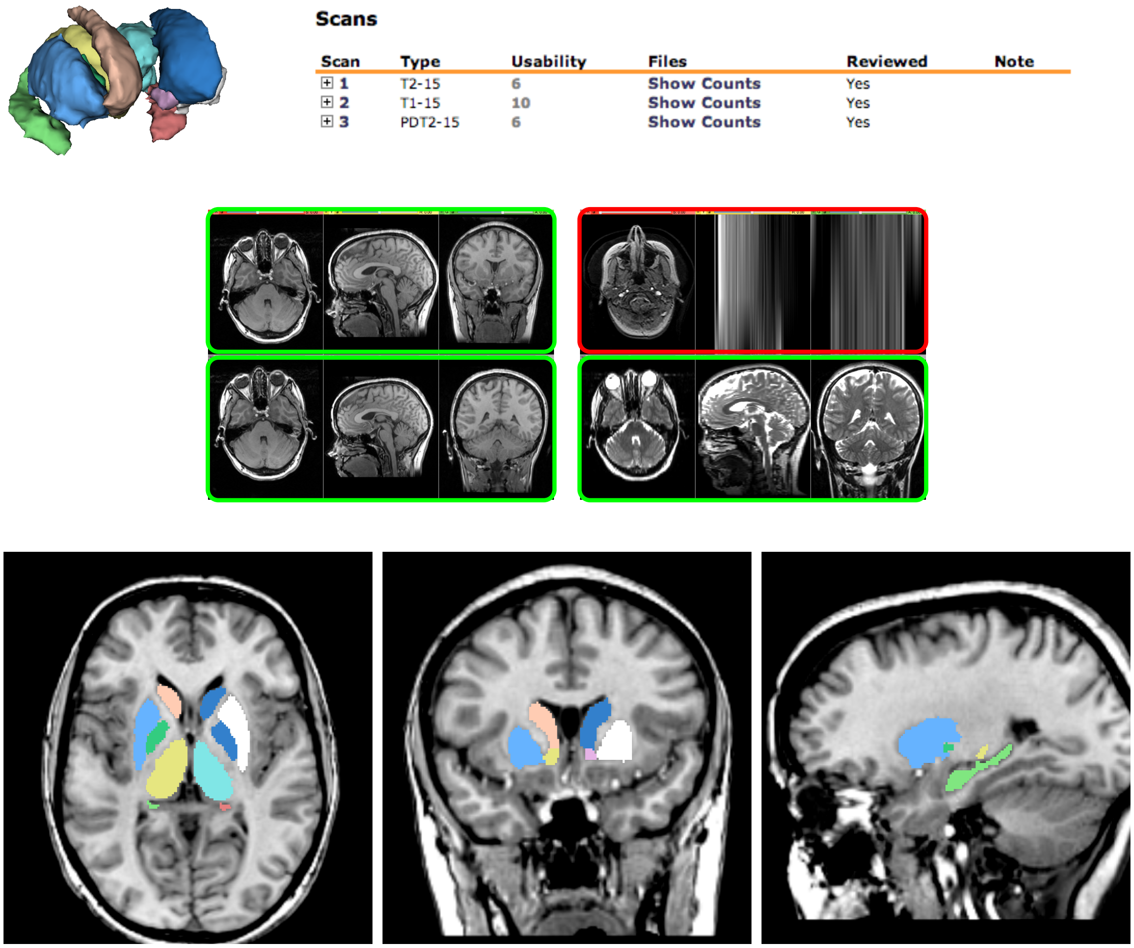
Figure A.8: A Good Subcortical segmentation examples from *BRAINSCut*. From the top left, **raw 1:** 3D volume rendered subcortical structure, a quality control scores for each raw scans, **raw 2:** a raw MRI examples scored above 5 (green) and under 5 (unusable, red) for the BAW processing, and **raw 3:** *BRAINSCut* results of six subcortical structures.



Scans

Scan	Type	Usability	Files	Reviewed	Note
⊞ 1	localizer	usable	Show Counts		
⊞ 2	localizer	usable	Show Counts		
⊞ 3	T1-15	6	Show Counts	Yes	
⊞ 4	PDT2-15	6	Show Counts	Yes	
⊞ 5	T1-15	0	Show Counts	Yes	Cropped and does not include entire head.

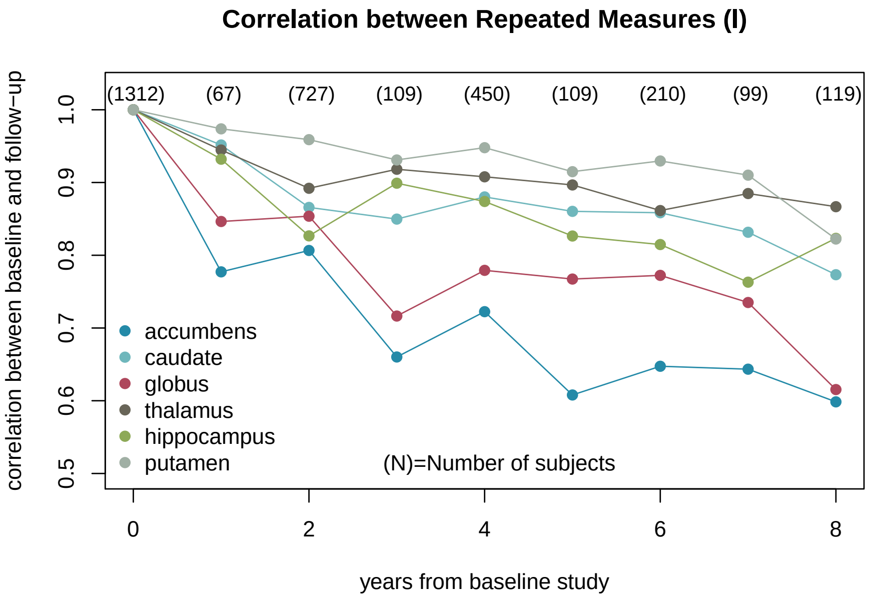
Figure A.9: A Bad Subcortical segmentation examples from *BRAINSCut*. From the top left, **raw 1:** 3D volume rendered subcortical structure, a quality control scores for each raw scans, **raw 2:** a raw MRI examples scored above 5 (green) and under 5 (unusable, red) for the BAW processing, and **raw 3:** *BRAINSCut* results of six subcortical structures.



Scans

Scan	Type	Usability	Files	Reviewed	Note
⊞ 1	T2-15	6	Show Counts	Yes	
⊞ 2	T1-15	10	Show Counts	Yes	
⊞ 3	PDT2-15	6	Show Counts	Yes	

Figure A.10: A Reasonable Subcortical segmentation examples from *BRAINSCut*. From the top left, **raw 1:** 3D volume rendered subcortical structure, a quality control scores for each raw scans, **raw 2:** a raw MRI examples scored above 5 (green) and under 5 (unusable, red) for the BAW processing, and **raw 3:** *BRAINSCut* results of six subcortical structures.



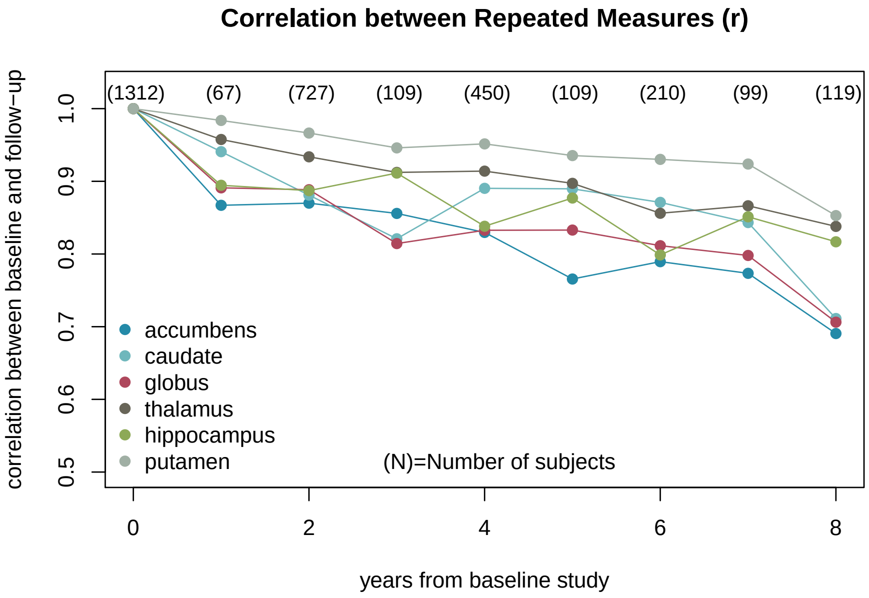


Figure A.11: Correlation between baseline and follow-up study for left (upper) and right (bottom) structures. Dicrease in correlation for farther time periods is very natural phenomena in longitudinal data

REFERENCES

[1] D.W. Aha and R.L. Bankert. A comparative evalution of sequential feature selection algorithms. *LECTURE NOTES IN STATISTICS-NEW YORK-SPRINGER VERLAG-*, pages 199–206, 1996.

[2] Ayelet Akselrod-Ballin, Meirav Galun, Moshe John Gomori, Ronen Basri, and Achi Brandt. Atlas guided identification of brain structures by combining 3D segmentation and SVM classification. *Medical Image Computing and Computer-Assisted Intervention*, 9(Pt 2):209–216, 2006.

[3] Pattern Analysis and Ricardo Gutierrez-osuna. L11 : sequential feature selection Feature extraction vs . feature selection. *Analysis*, pages 1–17.

[4] Petronella Anbeek, Koen L Vincken, Floris Groenendaal, Annemieke Koeman, Matthias J P van Osch, and Jeroen van der Grond. Probabilistic brain tissue segmentation in neonatal magnetic resonance imaging. *Pediatric research*, 63(2): 158–63, February 2008. ISSN 0031-3998. doi: 10.1203/PDR.0b013e31815ed071.

[5] A P Appelman, K L Vincken, Y van der Graaf, A L Vlek, T D Witkamp, W P Mali, M I Geerlings, A Algra, P A Doevendans, D E Grobbee, D E Rutten, L J Kappelle, F L Moll, and F L Visseren. White matter lesions and lacunar infarcts are independently and differently associated with brain atrophy: the SMART-MR study. *Cerebrovasc. Dis.*, 29(1):28–35, 2010.

[6] B B Avants, C L Epstein, M Grossman, and J C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, February 2008. ISSN 1361-8423. doi: 10.1016/j.media.2007.06.004.

[7] Brian B Avants, Nicholas J Tustison, Jue Wu, Philip a Cook, and James C Gee. An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics*, 9(4):381–400, December 2011. ISSN 1559-0089. doi: 10.1007/s12021-011-9109-y.

[8] Brian B Avants, Nicholas J Tustison, Jue Wu, Philip a Cook, and James C Gee. An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics*, 9(4):381–400, December 2011. ISSN 1559-0089. doi: 10.1007/s12021-011-9109-y.

[9] E H Aylward. Change in MRI striatal volumes as a biomarker in preclinical Huntington's disease. *Brain research bulletin*, 2007.

[10] E.H. Aylward, NB Anderson, FW Bylsma, MV Wagster, PE Barta, M. Sherr, J. Feeney, A. Davis, A. Rosenblatt, GD Pearlson, and Others. Frontal lobe volume in patients with Huntingtons disease. *Neurology*, 50(1):252–258, 1998.

[11] E.H Aylward, a Rosenblatt, K Field, V Yallapragada, K Kieburtz, M McDermott, L.a Raymond, E.W Almqvist, M Hayden, and C.a Ross. Caudate volume as an outcome measure in clinical trials for Huntingtons disease: a pilot study. *Brain Research Bulletin*, 62(2):137–141, December 2003. ISSN 03619230. doi: 10.1016/j.brainresbull.2003.09.005.

[12] Elizabeth H Aylward, Peggy C Nopoulos, Christopher A Ross, Douglas R Langbehn, Ronald K Pierson, James a Mills, Hans J Johnson, Vincent a Magnotta, Andrew R Juhl, Jane S Paulsen, PREDICT-HD Investigators and Coordinators of Huntington Study Group, and R Andrew. Longitudinal change in regional brain volumes in prodromal Huntington disease. *Journal of neurology, neurosurgery, and psychiatry*, 82(4):405–410, April 2011. ISSN 1468-330X. doi: 10.1136/jnnp.2010.208264.

[13] Elizabeth H Aylward, Dawei Liu, Peggy C Nopoulos, Christopher a Ross, Ronald K Pierson, James a Mills, Jeffrey D Long, and Jane S Paulsen. Striatal volume contributes to the prediction of onset of huntington disease in incident cases. *Biological psychiatry*, 71(9):822–8, May 2012. ISSN 1873-2402. doi: 10.1016/j.biopsych.2011.07.030.

[14] M. a. Balafar, a. R. Ramli, M. I. Saripan, and S. Mashohor. Review of brain MRI image segmentation methods. *Artificial Intelligence Review*, 33(3):261–274, January 2010. ISSN 0269-2821. doi: 10.1007/s10462-010-9155-0.

[15] M Bart and Haar Romeny. Multi-Scale and Multi-Orientation Medical Image Analysis. *Biomedical Image Processing*, pages 177–196, 2011. doi: 10.1007/978-3-642-15816-2.

[16] J T Becker, J Sanders, S K Madsen, A Ragin, L Kingsley, V Maruca, B Cohen, K Goodkin, E Martin, E N Miller, N Sacktor, J R Alger, P B Barker, P Saharan, O T Carmichael, P M Thompson, J B Margolick, H Armenian, B Crain, A Dobs, H Farzadegan, J Gallant, J Hylton, L Johnson, S Lai, O Selnes, J Shepard, C Thio, J P Phair, J S Chmiel, S Badri, C Conover, M O'Gorman, D Ostrow, F Palella, D Variakojis, S M Wolinsky, R Detels, B R Visscher, A Aronow, R Bolan, E Breen, A Butch, T Coates, R Effros, J Fahey, B Jamieson, O Martinez-Maza, J Oishi, P Satz, H Vinters, D Wiley, M Witt, O Yang, S Young, Z F Zhang, C R Rinaldo, L A Kingsley, R D Cranston, J J Martinson, J W Mellors, A J Silvestre, R D Stall, L P Jacobson, A Munoz, S R Cole, C Cox, G D'Souza, S J Gange, J Schollenberger, E C Seaberg, R E Huebner, G Dominguez, C McDonald, and P Brouwers. Subcortical brain atrophy persists even in HAART-regulated HIV disease. *Brain Imaging Behav*, 5 (2):77–85, June 2011.

[17] B Bilgic, A Bayram, A B Arslan, H Hanagasi, B Dursun, H Gurvit, M Emre, and E Lohmann. Differentiating symptomatic Parkin mutations carriers from

patients with idiopathic Parkinson's disease: contribution of automated segmentation neuroimaging method. *Parkinsonism Relat. Disord.*, 18(5):562–566, June 2012.

[18] K Blackmon, W B Barr, C Carlson, O Devinsky, J Dubois, D Pogash, B T Quinn, R Kuzniecky, E Halgren, and T Thesen. Structural evidence for involvement of a left amygdala-orbitofrontal network in subclinical anxiety. *Psychiatry Res*, 194(3):296–303, December 2011.

[19] C.A Bouman. Digital Image Processing: The visual Perception of Images. pages 1–22, 2013.

[20] Leo Breiman. Random forests. *Machine learning*, pages 1–33, 2001.

[21] Leo Breiman, Jerome Friedman, Charles Stone, and RA Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC; 1 edition, 1 edition, 1984. ISBN 0412048418.

[22] V a Cardenas, a T Du, D Hardin, F Ezekiel, P Weber, W J Jagust, H C Chui, N Schuff, and M W Weiner. Comparison of methods for measuring longitudinal brain change in cognitive impairment and dementia. *Neurobiology of aging*, 24 (4):537–44, 2003. ISSN 0197-4580.

[23] A Cerasa, A Quattrone, M C Gioia, A Magariello, M Muglia, F Assogna, S Bernardini, C Caltagirone, P Bossu, and G Spalletta. MAO A VNTR polymorphism and amygdala volume in healthy subjects. *Psychiatry Res*, 191(2): 87–91, February 2011.

[24] S E Chua, C Cheung, V Cheung, J T Tsang, E Y Chen, J C Wong, J P Cheung, L Yip, K S Tai, J Suckling, and G M McAlonan. Cerebral grey, white matter and csf in never-medicated, first-episode schizophrenia. *Schizophr. Res.*, 89 (1-3):12–21, January 2007.

[25] Andrea Ciarmiello, Milena Cannella, Secondo Lastoria, Maria Simonelli, Luigi Frati, David C Rubinsztein, and Ferdinando Squitieri. Brain White-Matter Volume Loss and Glucose Hypometabolism Precede the Clinical Symptoms of Huntington s Disease. *Journal of Nuclear Medicine*, L:215–222.

[26] U S Clark, R A Cohen, L H Sweet, A Gongvatana, K N Devlin, G N Hana, M L Westbrook, R C Mulligan, B A Jerskey, T L White, B Navia, and K T Tashima. Effects of HIV and early life stress on amygdala morphometry and neurocognitive function. *J Int Neuropsychol Soc*, 18(4):657–668, July 2012.

[27] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

[28] T O Dalaker, R Zivadinov, D P Ramasamy, M K Beyer, G Alves, K S Bronnick, O B Tysnes, D Aarsland, and J P Larsen. Ventricular enlargement and mild cognitive impairment in early Parkinson's disease. *Mov. Disord.*, 26(2):297–301, February 2011.

[29] Guorong Wu Hongjun Jia Daoqiang Zhang and Dinggang Shen. LNCS 6893 - Confidence-Guided Sequential Label Fusion for Multi-atlas Based Segmentation. pages 1–8, August 2011.

[30] R de Boer, M Schaap, F van der Lijn, H A Vrooman, M de Groot, A van der Lugt, M A Ikram, M W Vernooij, M M Breteler, and W J Niessen. Statistical analysis of minimum cost path based structural brain connectivity. *Neuroimage*, 55(2):557–565, March 2011.

[31] Renske de Boer, Henri a Vrooman, M Arfan Ikram, Meike W Vernooij, Monique M B Breteler, Aad van der Lugt, and Wiro J Niessen. Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. *NeuroImage*, 51(3):1047–56, July 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.03.012.

[32] J De Bresser, C Jongen, P Anbeek, M Viergever, L Kappelle, and G Biessels. Cross-sectional automatic measurement of brain volume on MRI: reproducibility of kNN-based probabilistic segmentation. In *Proceedings 17th Scientific Meeting, International Society for Magnetic Resonance in Medicine*, page 947, 2009.

[33] Jeroen de Bresser, Marileen P Portegies, Alexander Leemans, Geert Jan Biessels, L Jaap Kappelle, and Max a Viergever. A comparison of MR based segmentation methods for measuring brain atrophy progression. *NeuroImage*, 54(2): 760–8, January 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.09.060.

[34] L W de Jong, K van der Hiele, I M Veer, J J Houwing, R G Westendorp, E L Bollen, P W de Bruin, H A Middelkoop, M A van Buchem, and J van der Grond. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain*, 131(Pt 12):3277–3285, December 2008.

[35] B E Depue and M T Banich. Increased inhibition and enhancement of memory retrieval are associated with reduced hippocampal volume. *Hippocampus*, 22 (4):651–655, April 2012.

[36] C Derauf, B M Lester, N Neyzi, M Kekatpure, L Gracia, J Davis, K Kallianpur, J T Efird, and B Kosofsky. Subcortical and cortical structural central nervous system changes and attention processing deficits in preschool-aged children with prenatal methamphetamine and tobacco exposure. *Dev. Neurosci.*, 34(4):327–341, 2012.

[37] J Dewey, G Hana, T Russell, J Price, D McCaffrey, J Harezlak, E Sem, J C Anyanwu, C R Guttmann, B Navia, R Cohen, and D F Tate. Reliability and validity of MRI-based automated volumetry software relative to auto-assisted manual measurement of subcortical structures in HIV-infected patients from a multisite study. *Neuroimage*, 51(4):1334–1344, July 2010.

[38] Thomas G Dietterich. Ensemble Methods in Machine Learning. 1990.

[39] R A Dineen, C M Bradshaw, C S Constantinescu, and D P Auer. Extra-hippocampal subcortical limbic involvement predicts episodic recall performance in multiple sclerosis. *PLoS ONE*, 7(10):e44942, 2012.

[40] G Douaud, V Gaura, M-J Ribeiro, F Lethimonnier, R Maroy, C Verny, P Krystkowiak, P Damier, a C Bachoud-Levi, P Hantraye, and P Remy. Distribution of grey matter atrophy in Huntington's disease patients: a combined ROI-based and voxel-based morphometric study. *NeuroImage*, 32(4):1562–75, October 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.05.057.

[41] Marie-pierre Dubuisson, Ani E Jain, and East Lansing. A Modified Hausdorff Distance for Object Matching *. (1):566–568, 1994.

[42] C Eckerstrom, E Olsson, N Klasson, M Bjerke, M Gothlin, M Jonsson, S Rolstad, H Malmgren, A Wallin, and A Edman. High white matter lesion load is associated with hippocampal atrophy in mild cognitive impairment. *Dement Geriatr Cogn Disord*, 31(2):132–138, 2011.

[43] S. D. Edland, Y. Xu, M. Plevak, P. O'Brien, E. G. Tangalos, R. C. Petersen, and C. R. Jack. Total intracranial volume: Normative values and lack of association with Alzheimer's disease. *Neurology*, 59(2):272–274, July 2002. ISSN 0028-3878. doi: 10.1212/WNL.59.2.272.

[44] David M Erceg-Hurn and Vikki M Mirosevich. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *The American psychologist*, 63(7):591–601, October 2008. ISSN 0003-066X. doi: 10.1037/0003-066X.63.7.591.

[45] Zheng Fan, M Styner, J Muenzer, M Poe, and M Escolar. Correlation of automated volumetric analysis of brain MR imaging with cognitive impairment in a natural history study of mucopolysaccharidosis II. *AJNR. American journal of neuroradiology*, 31(7):1319–23, August 2010. ISSN 1936-959X. doi: 10.3174/ajnr.A2032.

[46] G Fein, V Di Sclafani, and J Tanabe. Hippocampal and cortical atrophy predict dementia in subcortical ischemic vascular disease. *Neurology*, 55(11):1626–1635, 2000.

[47] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre van der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, Albert Montillo, Nikos Makris, Bruce Rosen, and Anders M Dale. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–55, January 2002. ISSN 0896-6273.

[48] Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997. ISSN 00220000. doi: 10.1006/jcss.1997.1504.

[49] Luke Fu, Vladimir Fonov, Bruce Pike, Alan C Evans, and D Louis Collins. Automated analysis of multi site MRI phantom data for the NIHPD project. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 9(Pt 2):144–51, January 2006.

[50] M I Geerlings, A P Appelman, K L Vincken, A Algra, T D Witkamp, W P Mali, Y van der Graaf, P A Doevendans, D E Grobbee, G E Rutten, L J Kappelle, F L Moll, and F L Visseren. Brain volumes and cerebrovascular lesions on MRI in patients with atherosclerotic disease. The SMART-MR study. *Atherosclerosis*, 210(1):130–136, May 2010.

[51] Andrew Gelman, John B. Carlin, Hal S. Stern, and Conald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, second edi edition. ISBN 1420057294, 9781420057294.

[52] Emilie Gerardin, Gaël Chételat, Marie Chupin, Rémi Cuingnet, Béatrice Desgranges, Ho-Sung Kim, Marc Niethammer, Bruno Dubois, Stéphane Lehéricy, Line Garnero, Francis Eustache, and Olivier Colliot. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage*, 47(4):1476–1486, 2009.

[53] Charles J Geyer. Breakdown Point Theory Notes, 2006.

[54] Ali Ghayoor, Jatin G. Vaidya, and Hans J. Johnson. Development of a novel constellation based landmark detection algorithm. 8669:86693F–86693F–6, March 2013. doi: 10.1117/12.2006471.

[55] A L Goldman, L Pezawas, V S Mattay, B Fischl, B A Verchinski, B Zoltick, D R Weinberger, and A Meyer-Lindenberg. Heritability of brain morphology related to schizophrenia: a large-scale automated magnetic resonance imaging segmentation study. *Biol. Psychiatry*, 63(5):475–483, March 2008.

[56] Krzysztof Gorgolewski, Christopher D Burns, Cindee Madison, Dav Clark, Yaroslav O Halchenko, Michael L Waskom, and Satrajit S Ghosh. Nipype:

a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform*, 5, 08 2011. ISSN 1662-5196. doi: 10.3389/fninf.2011.00013.

[57] Ioannis S Gousias, Alexander Hammers, Serena J Counsell, Latha Srinivasan, Mary a Rutherford, Rolf a Heckemann, Jo V Hajnal, Daniel Rueckert, and a David Edwards. Magnetic resonance imaging of the newborn brain: automatic segmentation of brain images into 50 anatomical regions. *PloS one*, 8(4):e59990, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0059990.

[58] Sylvain Gouttard, Martin Styner, Marcel Prastawa, Joseph Piven, and Guido Gerig. Assessment of reliability of multi-site neuroimaging via traveling phantom study. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 11(Pt 2):263–70, January 2008.

[59] H Grydeland, K B Walhovd, L T Westlye, P Due-T?nnessen, V Ormaasen, ?. Sundseth, and A M Fjell. Amnesia following herpes simplex encephalitis: diffusion-tensor imaging uncovers reduced integrity of normal-appearing white matter. *Radiology*, 257(3):774–781, December 2010.

[60] Lei Guo, Xuena Liu, Youxi Wu, Weili Yan, and Xueqin Shen. Research on the segmentation of MRI image based on multi-classification support vector machine. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2007:6020–3, January 2007. ISSN 1557-170X. doi: 10.1109/IEMBS.2007.4353720.

[61] Lauren B Guthrie, Emily Oken, Jonathan a C Sterne, Matthew W Gillman, Rita Patel, Konstantin Vilchuck, Natalia Bogdanovich, Michael S Kramer, and Richard M Martin. Ongoing monitoring of data clustering in multicenter studies. *BMC medical research methodology*, 12(1):29, January 2012. ISSN 1471-2288. doi: 10.1186/1471-2288-12-29.

[62] Mark Hall, Hazeltine National, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA Data Mining Software : An Update. 11(1):10–18.

[63] Mark A Hall. Correlation-based Feature Selection for Machine Learning. (April), 1999.

[64] Frank Hampel. Robust statistics : A brief introduction and overview. 2001.

[65] Frank R. Hampel, Elvezio M. Ronchetti, Pter J. Rousseeuw, and Werner A. Stahel. *Robust statistics: the approach based on influence functions*, volume 29. Wiley, May 1986. doi: 10.1080/00401706.1987.10488218.

[66] Gordon J Harris, Godfrey D Pearlson, Carol E Peyser, Elizabeth H Aylward, Joy Roberts, Patrick E Barta, Gary A Chase, and Susan E Folstein. Putamen Volume Reduction on Magnetic Resonance Imaging Exceeds Caudate Changes in Mild Huntington s Disease. *Annals of Neurology*, 31(1):69–75, 1992.

[67] T Hartley and R Harlow. An association between human hippocampal volume and topographical memory in healthy young adults. *Front Hum Neurosci*, 6: 338, 2012.

[68] Rolf a Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–26, October 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.05.061.

[69] Joseph P Hornak. The Basics of MRI. *Biomedical Engineering*, 24(2003):2–6, 2008.

[70] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[71] W Iba and P Langley. Induction of One-Level Decision Trees. *Proceedings of the Ninth International Conference on . . .*, 1992.

[72] G Iordanescu, P Venkatasubramanian, and A Wyrwicz. Segmentation of Amyloid Plaques in MR images of the APP Transgenic Mouse Brain using SVM. In *Proceedings 17th Scientific Meeting, International Society for Magnetic Resonance in Medicine*, page 2871, 2009.

[73] N Japkowicz. The class imbalance problem: a systematic study. *Intelligent data analysis*, 2002.

[74] H M Jochemsen, M Muller, F L Visseren, P Scheltens, K L Vincken, W P Mali, Y van der Graaf, and M I Geerlings. Blood Pressure and Progression of Brain Atrophy: The SMART-MR Study. *JAMA Neurol*, 70(8):1046–1053, August 2013.

[75] Pierre-Marc Jodoin, Max Mignotte, and Christophe Rosenberger. Segmentation framework based on label field fusion. *IEEE Transactions on Image Processing*, 16(10):2535–2550, 2007.

[76] J Jovicich, M Marizzoni, R Sala-Llonch, B Bosch, D Bartres-Faz, J Arnold, J Benninghoff, J Wiltfang, L Roccatagliata, F Nobili, T Hensch, A Trankner, P Schonknecht, M Leroy, R Lopes, R Bordet, V Chanoine, J P Ranjeva, M Didic, H Gros-Dagnac, P Payoux, G Zoccatelli, F Alessandrini, A Beltramello, N Bargallo, O Blin, and G B Frisoni. Brain morphometry reproducibility in multi-center 3T MRI studies: A comparison of cross-sectional and longitudinal segmentations. *Neuroimage*, 83C:472–484, May 2013.

[77] Jorge Jovicich, Silvester Czanner, Douglas Greve, Elizabeth Haley, Andre van der Kouwe, Randy Gollub, David Kennedy, Franz Schmitt, Gregory Brown, James Macfall, Bruce Fischl, and Anders Dale. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30(2):436–43, April 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2005.09.046.

[78] Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deborah Blacker, Paul Maguire, Diana Rosas, Nikos Makris, Randy Gollub, Anders Dale, Bradford C Dickerson, and Bruce Fischl. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage*, 46(1):177–92, May 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.02.010.

[79] Monica Juuhl-Langseth, Lars M Rimol, Inge A Rasmussen, Rune Thormodsen, Aina Holmén, Kyrre E Emblem, Paulina Due-Tø nnessen, Bjørn Rishovd Rund, Ingrid Agartz, A Holmen, and P Due-T?nnessen. Comprehensive segmentation of subcortical brain volumes in early onset schizophrenia reveals limited structural abnormalities. *Psychiatry Res*, 203(1):14–23, July 2012. ISSN 1872-7123. doi: 10.1016/j.pscychresns.2011.10.005.

[80] Matthew J Kempton, Tracy S a Underwood, Simon Brunton, Floris Stylios, Anne Schmechtig, Ulrich Ettinger, Marcus S Smith, Simon Lovestone, William R Crum, Sophia Frangou, Steven C R Williams, and Andrew Simmons. A comprehensive testing protocol for MRI neuroanatomical segmentation techniques: Evaluation of a novel lateral ventricle segmentation method. *NeuroImage*, 58(4):1051–9, October 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.06.080.

[81] A R Khan and M F Beg. Multi-structure whole brain registration and population average. *Conf Proc IEEE Eng Med Biol Soc*, 2009:5797–5800, 2009.

[82] E Y Regina Kim, Hans J Johnson, and Norman K Williams. Affine Transformation for Landmark Based Registration Initializer in ITK. *Engineering*, pages 1–8, 2011.

[83] Eun Young Kim. ACADEMIC CERTIFICATION FOR :. page 52242, 2013.

[84] Eun Young Kim and Hans Johnson. Multi-structure segmentation of multimodal brain images using artificial neural networks. *Analysis*, 7623:76234B–76234B–12, 2010. doi: 10.1117/12.844613.

[85] Carole L Kimberlin and Almut G Winterstein. Validity and reliability of measurement instruments used in research. *American journal of health-*

system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists, 65(23):2276–84, December 2008. ISSN 1535-2900. doi: 10.2146/ajhp070364.

[86] Daphne Koller and Gates Building. Toward Optimal Feature Selection.

[87] C H Lai and Y T Wu. Duloxetine's modest short-term influences in subcortical structures of first episode drug-naïve patients with major depressive disorder and panic disorder. *Psychiatry Res*, 194(2):157–162, November 2011.

[88] Kenneth Ivan Laws. Textured Image Segmentation. January 1980.

[89] Guy Lebanon. Bias , Variance , and MSE of Estimators. Technical Report 1, 2010.

[90] Sang Hak Lee, Hyung Il Koo, and Nam Ik Cho. Image segmentation algorithms based on the machine learning of features. *Pattern Recognition Letters*, 31(14): 2325–2336, 2010. ISSN 01678655. doi: 10.1016/j.patrec.2010.07.004.

[91] Victor Lempitsky and Michael Verhoek. Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. *Functional Imaging and . . .*, pages 447–456, 2009.

[92] Jeffery D Long, Jane S. Paulsen, Karen Marder, Ying Zhang, Ji-In Kim, and James a Mills. Tracking motor impairments in the progression of Huntington's disease. *Movement disorders : official journal of the Movement Disorder Society*, 00(00):1–9, October 2013. ISSN 1531-8257. doi: 10.1002/mds.25657.

[93] X Long, W Liao, C Jiang, D Liang, B Qiu, and L Zhang. Healthy aging: an automatic analysis of global and regional morphological alterations of human brain. *Acad Radiol*, 19(7):785–793, July 2012.

[94] Vincent A. Magnotta, Joy Tamiko Matsui, Dawei Liu, Hans J. Johnson, Jeffrey D Long, Bradley D Bolster Jr, Byron A. Mueller, Kelvin O. Lim, Susumu Mori, Karl G Helmer, and Others. Multi-Center Reliability of Diffusion Tensor Imaging. *Brain . . .*, 2(6):345–355, 2012.

[95] D S Majid, A R Aron, W Thompson, S Sheldon, S Hamza, D Stoffers, D Holland, J Goldstein, J Corey-Bloom, and A M Dale. Basal ganglia atrophy in prodromal Huntington's disease is detectable over one year using automated segmentation. *Mov. Disord.*, 26(14):2544–2551, December 2011.

[96] M Mallar Chakravarty, P Steadman, M C van Eede, R D Calcott, V Gu, P Shaw, A Raznahan, D Louis Collins, and J P Lerch. Performing label-fusion-based segmentation using multiple automatically generated templates. *Hum Brain Mapp*, May 2012.

[97] Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996. ISSN 1082-989X. doi: 10.1037//1082-989X.1.1.30.

[98] R A Menke, K Szewczyk-Krolikowski, S Jbabdi, M Jenkinson, K Talbot, C E Mackay, and M Hu. Comprehensive morphometry of subcortical grey matter structures in early-stage Parkinson's disease. *Hum Brain Mapp*, July 2013.

[99] Demetrio Messina, Antonio Cerasa, Francesca Condino, Gennarina Arabia, Fabiana Novellino, Giuseppe Nicoletti, Maria Salsone, Maurizio Morelli, Pier Luigi Lanza, and Aldo Quattrone. Patterns of brain atrophy in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Parkinsonism {&} related disorders*, 17(3):172–176, March 2011.

[100] Maria Carolina Monard and Gustavo E A P A Batista. Learning with Skewed Class Distributions. *Learning*, pages 1–9, 2003.

[101] Alonso Montoya, Bruce H Price, Matthew Menear, and Martin Lepage. Examen critique Brain imaging and cognitive dysfunctions in Huntington s disease. *Imaging*, 31(1):21–29, 2006.

[102] Rajendra a Morey, Elizabeth S Selgrade, Henry Ryan Wagner, Scott a Huettel, Lihong Wang, and Gregory McCarthy. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Human brain mapping*, 31(11):1751–62, November 2010. ISSN 1097-0193. doi: 10.1002/hbm.20973.

[103] J. Morra, Z. Tu, A. Toga, and P. Thompson. Lossless Online Ensemble Learning (LOEL) and Its Application to Subcortical Segmentation. *Medical Image Computing and Computer-Assisted InterventionMICCAI 2009*, pages 432–440, 2009.

[104] Jonathan H Morra, Zhuowen Tu, Liana G Apostolova, Amity E Green, Arthur W Toga, and Paul M Thompson. Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. *IEEE Transactions on Medical Imaging*, 29(1):30–43, 2010.

[105] M Muller, Y van der Graaf, A Algra, J Hendrikse, W P Mali, M I Geerlings, P Doevendans, D Grobbee, G Rutten, L Kappelle, W Mali, F Moll, and F Visseren. Carotid atherosclerosis and progression of brain atrophy: the SMART-MR study. *Ann. Neurol.*, 70(2):237–244, August 2011.

[106] D Mungas, D Harvey, B R Reed, W J Jagust, C DeCarli, L Beckett, W J Mack, J H Kramer, M W Weiner, N Schuff, and H C Chui. Longitudinal volumetric MRI change and rate of cognitive decline. *Neurology*, 65(4):565–71, August 2005. ISSN 1526-632X. doi: 10.1212/01.wnl.0000172913.88973.0d.

[107] PC Nopoulos, EH Aylward, and CA Ross. Cerebral cortex structure in prodromal Huntington disease. *Neurobiology of diseaseof Disease*, 40:544–554, 2010.

[108] Peggy C Nopoulos, Elizabeth H Aylward, Christopher A Ross, James A Mills, Douglas R Langbehn, Hans J Johnson, Vincent A Magnotta, Ronald K Pierson, Leigh J Beglinger, Martha A Nance, Roger A Barker, Jane S Paulsen, and PREDICT-HD Investigators and Coordinators of the Huntington Study Group. Smaller intracranial volume in prodromal Huntington's disease: evidence for abnormal neurodevelopment. *Brain : a journal of neurology*, 134(Pt 1):137–142, January 2011.

[109] Y Ostby, C K Tamnes, A M Fjell, L T Westlye, P Due-T?nnessen, and K B Walhovd. Heterogeneity in subcortical brain development: A structural magnetic resonance imaging study of brain maturation from 8 to 30 years. *J. Neurosci.*, 29(38):11772–11782, September 2009.

[110] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining* . Addison Wesley; 1 edition, 2005. ISBN 0321321367.

[111] Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56(3):907–922, June 2011.

[112] R H Paul, A M Brickman, B Navia, C Hinkin, P F Malloy, A L Jefferson, R A Cohen, D F Tate, and T P Flanigan. Apathy is associated with volume of the nucleus accumbens in patients infected with HIV. *J Neuropsychiatry Clin Neurosci*, 17(2):167–171, 2005.

[113] Jane S Paulsen, Vince a Magnotta, Ania E Mikos, Henry L Paulson, Elizabeth Penziner, Nancy C Andreasen, and Peg C Nopoulos. Brain structure in preclinical Huntington's disease. *Biological psychiatry*, 59(1):57–63, January 2006. ISSN 0006-3223. doi: 10.1016/j.biopsych.2005.06.003.

[114] JS Paulsen, PC Nopoulos, E Aylward, and CA Ross. Striatal and white matter predictors of estimated diagnosis for Huntington disease. *Brain research*, 2010.

[115] JS S Paulsen, DR R Langbehn, JC C Stout, Langbehn, E Aylward, CA a Ross, M Nance, M Guttman, S Johnson, M MacDonald, LJ J Beglinger, K Duff, E Kayson, K Biglan, I Shoulson, D Oakes, M Hayden, and Iacothsg Predict-HD. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *Journal of neurology, neurosurgery, and psychiatry*, 79(8):874–80, August 2008. ISSN 1468-330X. doi: 10.1136/jnnp.2007.128728.

[116] Frank Perbet. Random Forest Clustering and Application to Video Segmentation. *Science*, (d):1–10, 2008.

[117] P Perona. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence,*, 1990.

[118] Maria Petrou, Agma J M Traina, Caetano Traina Jr, Marcela X Ribeiro, Pedro H Bugatti, Carolina Y V Watanabe, Paulo M Azevedo-marques, M Bart, and Haar Romeny. *Biomedical Image Processing*. Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-15815-5. doi: 10.1007/978-3-642-15816-2.

[119] Ronald Pierson, Hans Johnson, Gregory Harris, Helen Keefe, Jane S Paulsen, Nancy C Andreasen, and Vincent A Magnotta. Fully automated analysis using BRAINS: AutoWorkup. *NeuroImage*, 54(1):328–336, January 2011.

[120] Robert a Pooley. AAPM/RSNA physics tutorial for residents: fundamental physics of MR imaging. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 25(4):1087–99, 2005. ISSN 1527-1323. doi: 10.1148/rg.254055027.

[121] Stephanie Powell. Automated brain segmentation using neural networks. In *Medical Imaging 2006: Image Processing*, pages 61443Q—-61443Q—-11. SPIE, 2006.

[122] Stephanie Powell, Vincent a Magnotta, Hans Johnson, Vamsi K Jammalamadaka, Ronald Pierson, and Nancy C Andreasen. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. *NeuroImage*, 39(1):238–47, January 2008. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.05.063.

[123] Marcel Prastawa, Elizabeth Bullitt, Nathan Moon, Koen Van Leemput, and Guido Gerig. Automatic brain tumor segmentation by subject specific modification of atlas priors. *Academic Radiology*, 10(12):1341–1348, December 2003. ISSN 10766332. doi: 10.1016/S1076-6332(03)00506-3.

[124] Marcel Prastawa, John Gilmore, W Lin, and G Gerig. Automatic segmentation of neonatal brain MRI. *Medical Image Computing and Computer-Assisted Intervention*, pages 10–17, 2004.

[125] Marcel Prastawa, John H Gilmore, Weili Lin, and Guido Gerig. Automatic segmentation of MR images of the developing newborn brain. *Medical image analysis*, 9(5):457–66, October 2005. ISSN 1361-8415. doi: 10.1016/j.media.2005.05.007.

[126] Marcelinus Prastawa. *An MRI Segmentation Framework for Brains with Anatomical Deviations*. PhD thesis, University of Chapel Hill, 2007.

[127] predict-hd. Predict-hd offical site: https://www.predict-hd.net/.

[128] F Provost, T Fawcett, and R Kohavi. The case against accuracy estimation for comparing induction algorithms. *Proceedings of the fifteenth ...*, 1998.

[129] Azhar Quddus, Paul Fieguth, and Otman Basir. Adaboost and Support Vector Machines for White Matter Lesion Segmentation in MR Images. *Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, 1:463–466, 2005.

[130] D P Ramasamy, R H Benedict, J L Cox, D Fritz, N Abdelrahman, S Hussein, A Minagar, M G Dwyer, and R Zivadinov. Extent of cerebellum, subcortical and cortical atrophy in patients with MS: a case-control study. *J. Neurol. Sci.*, 282(1-2):47–54, July 2009.

[131] K B Ramesh. Linear Feature Extraction. *Computer Graphics and Image Processing*, pages 257–269, 1980.

[132] K P Rankin, H J Rosen, J H Kramer, G F Schauer, M W Weiner, N Schuff, and B L Miller. Right and left medial orbitofrontal volumes show an opposite relationship to agreeableness in FTD. *Dement Geriatr Cogn Disord*, 17(4): 328–332, 2004.

[133] Xiaofeng Ren. A probabilistic multi-scale model for contour completion based on image statistics. *Science*, 2002.

[134] Martin Reuter and Bruce Fischl. Avoiding asymmetry-induced bias in longitudinal image processing. *NeuroImage*, 57(1):19–21, July 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.02.076.

[135] Elaine Rich and Kevin Knight. *Artificial intelligence*. McGraw-Hill, 2 edition, 1991. ISBN 0070522634, 9780070522633.

[136] B.D Ripley. Robust statistics. pages 1–11, 1992.

[137] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2): 1–39, November 2009. ISSN 0269-2821. doi: 10.1007/s10462-009-9124-7.

[138] Su Ruan and Stephane Lebonvallet. Multi-kernel SVM based classification for brain tumor segmentation of MRI multi-sequence. *2009 16th IEEE International Conference on Image Processing ICIP*, pages 3373–3376, 2009. doi: 10.1109/ICIP.2009.5413878.

[139] RM Rubin. Principles of imaging in neuro-ophthalmology. *Ophthalmology. Mosby, Philadelphia*, pages 943–949, 1999. doi: 10.1016/B978-0-323-04332-8.00156-6.

[140] Mert R Sabuncu, B T Thomas Yeo, Koen Van Leemput, Bruce Fischl, and Polina Golland. A generative model for image segmentation based on label fusion. *IEEE transactions on medical imaging*, 29(10):1714–29, October 2010. ISSN 1558-254X. doi: 10.1109/TMI.2010.2050897.

[141] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Object Class Segmentation using Random Forests. *Engineering*, pages 1–8, 2008.

[142] D L Schwartz, A D Mitchell, D L Lahna, H S Luber, M S Huckans, S H Mitchell, and W F Hoffman. Global and local morphometric differences in recently abstinent methamphetamine-dependent individuals. *Neuroimage*, 50 (4):1392–1401, May 2010.

[143] Neeraj Sharma, Amit K Ray, Shiru Sharma, K K Shukla, Satyajit Pradhan, and Lalit M Aggarwal. Segmentation and classification of medical images using texture-primitive features: Application of BAM-type artificial neural network. *Journal of medical physics Association of Medical Physicists of India*, 33(3): 119–126, 2008.

[144] N Shiee, P L Bazin, K M Zackowski, S K Farrell, D M Harrison, S D Newsome, J N Ratchford, B S Caffo, P A Calabresi, D L Pham, and D S Reich. Revisiting brain atrophy and its relationship to disability in multiple sclerosis. *PLoS ONE*, 7(5):e37049, 2012.

[145] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

[146] E E Smith, D H Salat, J Jeng, C R McCreary, B Fischl, J D Schmahmann, B C Dickerson, A Viswanathan, M S Albert, D Blacker, and S M Greenberg. Correlations between MRI white matter lesion location and executive function and episodic memory. *Neurology*, 76(17):1492–1499, April 2011.

[147] Travis B Smith and Krishna S Nayak. MRI artifacts and correction strategies. *Imaging in Medicine*, 2(4):445–457, August 2010. ISSN 1755-5191. doi: 10.2217/iim.10.33.

[148] Irwin Sobel. An isotropic 3x3x3 volume gradient operator. *Hewlett-Packard Laboratories*, 1995.

[149] C. Soneson, M. Fontes, Y. Zhou, V. Denisov, J.S. Paulsen, D. Kirik, Å. Petersén, and Others. Early changes in the hypothalamic region in prodromal Huntington disease revealed by MRI analysis. *Neurobiology of disease*, 40(3):531–543, 2010.

[150] H Späth. Fitting affine and orthogonal transformations between two sets of points. *Mathematical Communications*, 2004.

[151] I Spoletini, F Piras, S Fagioli, I A Rubino, G Martinotti, A Siracusano, C Caltagirone, and G Spalletta. Suicidal attempts and increased right amygdala volume in schizophrenia. *Schizophr. Res.*, 125(1):30–40, January 2011.

[152] Sheila Sprague, Joel M Matta, Mohit Bhandari, David Dodgin, Charles R Clark, Phil Kregor, Gary Bradley, and Lester Little. Multicenter collaboration in

observational research: improving generalizability and efficiency. *The Journal of bone and joint surgery. American volume*, 91 Suppl 3:80–6, May 2009. ISSN 1535-1386. doi: 10.2106/JBJS.H.01623.

[153] S E Starkstein, F Bylsma, C I Peyser, M Folstein, and S E Folstein. Neuroradiology Neuropsychological correlates of brain atrophy in Huntington ' s disease : a magnetic resonance imaging study. *Magnetic Resonance Imaging*, pages 487–489, 1992.

[154] E V Sullivan, A Pfefferbaum, T Rohlfing, F C Baker, M L Padilla, and I M Colrain. Developmental change in regional brain structure over 7 months in early adolescence: comparison of approaches for longitudinal atlas-based parcellation. *Neuroimage*, 57(1):214–224, July 2011.

[155] Motofumi T Suzuki, Yoshitomo Yaginuma, Tsuneo Yamada, and Yasutaka Shimizu. A Shape Feature Extraction Method Based on 3D Convolution Masks. *Symposium A Quarterly Journal In Modern Foreign Literatures*, 2006.

[156] Sarah J Tabrizi, Douglas R Langbehn, Blair R Leavitt, Raymund Ac Roos, Alexandra Durr, David Craufurd, Christopher Kennard, Stephen L Hicks, Nick C Fox, Rachael I Scahill, Beth Borowsky, Allan J Tobin, H Diana Rosas, Hans Johnson, Ralf Reilmann, Bernhard Landwehrmeyer, and Julie C Stout. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet neurology*, 8(9):791–801, September 2009. ISSN 1474-4422. doi: 10.1016/S1474-4422(09) 70170-X.

[157] I-F. Talos, M. Jakab, R. Kikinis, and M.E.. Shenton. Spl-pnl brain atlas. 03 2008.

[158] S J Teipel, G E Alexander, M B Schapiro, H J Moller, S I Rapoport, and H Hampel. Age-related cortical grey matter reductions in non-demented Down's syndrome adults determined by MRI with voxel-based morphometry. *Brain*, 127 (Pt 4):811–824, April 2004.

[159] S Tinaz, M G Courtney, and C E Stern. Focal cortical and subcortical atrophy in early Parkinson's disease. *Mov. Disord.*, 26(3):436–441, February 2011.

[160] F Torelli, N Moscufo, G Garreffa, F Placidi, A Romigi, S Zannino, M Bozzali, F Fasano, G Giulietti, I Djonlagic, A Malhotra, M G Marciani, and C R Guttmann. Cognitive profile and brain morphological changes in obstructive sleep apnea. *Neuroimage*, 54(2):787–793, January 2011.

[161] trackon. Track-on.

[162] A D Turner, M L Furey, W C Drevets, C Zarate, and A C Nugent. Association between subcortical volumes and verbal memory in unmedicated depressed patients and healthy controls. *Neuropsychologia*, 50(9):2348–2355, July 2012.

[163] M Vaidyanathan, L P Clarke, C Heidtman, R P Velthuizen, and L O Hall. Normal brain volume measurements using multispectral MRI segmentation. *Magnetic Resonance Imaging*, 15(1):87–97, 1997.

[164] P O Valko, J Hanggi, M Meyer, and H H Jung. Evolution of striatal degeneration in McLeod syndrome. *Eur. J. Neurol.*, 17(4):612–618, April 2010.

[165] S J van den Bogaard, E M Dumas, L Ferrarini, J Milles, M A van Buchem, J van der Grond, and R A Roos. Shape analysis of subcortical nuclei in Huntington's disease, global versus local atrophy–results from the TRACK-HD study. *J. Neurol. Sci.*, 307(1-2):60–68, August 2011.

[166] K Van Leemput, F Maes, D Vandermeulen, and P Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE transactions on medical imaging*, 18(10):885–96, October 1999. ISSN 0278-0062. doi: 10.1109/42.811268.

[167] G T Vasileiadis, N Gelman, V K Han, L A Williams, R Mann, Y Bureau, and R T Thompson. Uncomplicated intraventricular hemorrhage is followed by reduced cortical volume at near-term age. *Pediatrics*, 114(3):e367—-372, September 2004.

[168] Henri A Vrooman, Chris A Cocosco, Fedde Van Der Lijn, Rik Stokking, M Arfan Ikram, Meike W Vernooij, Monique M B Breteler, and Wiro J Niessen. Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. *NeuroImage*, 37(1):71–81, 2007.

[169] Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John Pluta, Caryne Craige, and Paul a Yushkevich. Multi-Atlas Segmentation with Joint Label Fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623, June 2012. ISSN 1939-3539. doi: 10.1109/TPAMI.2012.143.

[170] Hongzhi Wang, Paul Yushkevich, and Paul A Yushkevich. Multi-Atlas Segmentation with Joint Label Fusion and Corrective Learning - An Open Source Implementation Frontiers in Bioengineering Multi-Atlas Segmentation with Joint Label Fusion and Corrective Learning - An Open Source Implementation. 2013.

[171] C Watson, F Andermann, P Gloor, M Jones-Gotman, T Peters, a Evans, a Olivier, D Melanson, and G Leroux. Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. *Neurology*, 42(9):1743–50, September 1992. ISSN 0028-3878.

[172] W M Wells, W L Grimson, R Kikinis, and F a Jolesz. Adaptive segmentation of MRI data. *IEEE transactions on medical imaging*, 15(4):429–42, January 1996. ISSN 0278-0062. doi: 10.1109/42.511747.

[173] R L Widya, A de Roos, S Trompet, A J de Craen, R G Westendorp, J W Smit, M A van Buchem, J van der Grond, J Shepherd, S M Cobbe, I Ford, A Gaw, P W Macfarlane, C J Packard, D J Stott, G J Blauw, E Bollen, A M Kamper, M B Murphy, B M Buckley, M Hyland, I J Perry, J W Jukema, A E Meinders, B J Sweeney, C Twomey, H C Diener, J Feely, T Pearson, S Pocock, and P van Zwieten. Increased amygdalar and hippocampal volumes in elderly obese individuals with or at risk of cardiovascular disease. *Am. J. Clin. Nutr.*, 93(6):1190–1195, June 2011.

[174] R C Wolf, P A Thomann, A K Thomann, N Vasic, N D Wolf, G B Landwehrmeyer, and M Orth. Brain structure in preclinical Huntington's disease: a multi-method approach. *Neurodegener Dis*, 12(1):13–22, 2013.

[175] J S Wonderlick, D a Ziegler, P Hosseini-Varnamkhasti, J J Locascio, a Bakkour, a van der Kouwe, C Triantafyllou, S Corkin, and B C Dickerson. Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. *NeuroImage*, 44(4):1324–33, February 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2008.10.037.

[176] Dong Xu and Hua Li. Geometric moment invariants. *Pattern Recognition*, 41 (1):240–249, January 2008. ISSN 00313203. doi: 10.1016/j.patcog.2007.05.001.

[177] Zhao Yi, Antonio Criminisi, Jamie Shotton, and Andrew Blake. Discriminative, semantic segmentation of brain tissue in MR images. *Medical Image Computing and Computer-Assisted Intervention*, 12(Pt 2):558–565, 2009.

[178] Bofeng Zhang, Wenhao Zhu, Hui Zhu, Anping Song, and Wu Zhang. A SVM based automatic segmentation method for brain magnetic resonance image series. In *Symposia and Workshops Held in Conjunction with the 7th International Conference on Ubiquitous Intelligence and Computing UIC 2010 and the 7th International Conference on Autonomic and Trusted Computing ATC 2010 October 26 2010 October 29 2010*, pages 375–379. IEEE Computer Society, 2010.

[179] Lei Zhang, Xun Wang, Nicholas Penwarden, and Qiang Ji. An Image Segmentation Framework Based on Patch Segmentation Fusion. In *EUSIPCO 2006*, volume 00, pages 1–5. Ieee, 2006. ISBN 0769525210. doi: 10.1109/ICPR.2006.250.

[180] Ying Zhang, Jeffrey D Long, James a Mills, John H Warner, Wenjing Lu, and Jane S Paulsen. Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 156B(7):751–63, December 2011. ISSN 1552-485X. doi: 10.1002/ajmg.b.31232.

[181] J Zhou, K L Chan, V F Chong, and S M Krishnan. Extraction of brain tumor from MR images using one-class support vector machine. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 6:6411–4, January 2005. ISSN 1557-170X. doi: 10.1109/IEMBS.2005.1615965.

[182] Alex P Zijdenbos, Reza Forghani, and Alan C Evans. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE transactions on medical imaging*, 21(10):1280–91, October 2002. ISSN 0278-0062. doi: 10.1109/TMI.2002.806283.