MRI Brain Analysis Testbed (BAT): Methodology and Automatic Validation Pipeline

Oleg Ivanov

A Master of Science Thesis

Department of Biomedical Engineering

McGill University

Montreal, Quebec, Canada

August 2005

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Engineering

Your file Votre référence
ISBN: 978-0-494-24970-3
Our file Notre référence
ISBN: 978-0-494-24970-3

ABSTRACT

Magnetic Resonance Imaging (MRI) is extensively used in brain imaging research and clinical diagnostics. Increasingly, automated image processing algorithms are used for identification of tissue types within the image, such as gray matter, white matter and cerebro-spinal fluid. There is a wide range of algorithms, which vary in speed and accuracy, and it is often difficult to compare their performance in any objective and controlled fashion. The goal of this research was to design an automatic, generic, standard, extensible pipeline for objective and quantitative validation of MRI tissue classification algorithms and their processing pipelines. The main issues and requirements, for objective validation of different algorithms, are the use of common terminology, methodology, standard validation data sets, corresponding ground truth, validation metrics and statistical foundation. Based on those requirements, an automatic Brain Analysis Testbed (BAT) was developed to determine an objective evaluation score for MRI processing method. BAT supports Montreal Neurological Institute on-site or off-site processing of MRI data, accessible by a web interface (http://www.bic.mni.mcgill.ca/validation/). Validation results are stored in the BAT database permanently, allowing the comparison of newly developed processing methods with existing ones. Furthermore, BAT can be used to determine the optimal classification parameters, or the best classifier algorithm for a specific MRI classification purpose, simply by searching the BAT database. The main purposes and principles of BAT are demonstrated with some practical MRI processing examples.

RÉSUMÉ

L'Imagerie par Résonance Magnétique (IRM) est largement utilisée à des fins de recherche en imagerie cérébrale ou de diagnostiques cliniques. Les algorithmes automatisés de traitement d'images sont de plus en plus utilisés pour l'identification de différents types de tissus dans les images, tels que la matière grise, la matière blanche et le liquide cérébrospinal. Il y a une grande variété d'algorithmes qui varient dans la vitesse et la précision et il est souvent difficile de comparer leur performance d'une façon objective et contrôlée. Le but de cette recherche était de concevoir un pipeline automatique, générique, standard et extensible afin de permettre une validation quantitative d'algorithmes de classification de tissu d'IRM et leurs pipelines de traitement. Les conditions principales pour la validation objective d'algorithmes différents sont : l'usage d'une terminologie et méthodologie commune; les validations des données standards; la vérification des algorithmes avec des données simulées, des métriques de validation et des appuis statistiques. A partir de ces conditions, un « Brain Analysis Testbed » (BAT) a été développé pour obtenir un résultat d'évaluation objectif pour chacune des méthodes de traitement d'IRM. BAT permet le traitement des données d'IRM ou sur place à l'Institut Neurologique de Montréal ou sur l'interface en ligne (http://www.bic.mni.mcgill.ca/validation/). Les résultats de validation sont enregistrés dans la base de données de BAT, permettant d'une façon permanente la comparaison des méthodes récemment développées avec les méthodes déjà validées. De plus, BAT peut être utilisé pour déterminer les paramètres de classification optimaux, ou le meilleur algorithme de classificateur pour un but spécifique de classification d'IRM, et ce, seulement en cherchant dans la base de données de BAT. Les objectifs généraux et spécifiques de BAT ont été testés avec quelques exemples de traitement d'IRM.

ACKNOWLEDGMENTS

Above all, I would like to thank my supervisor Dr. Alan C. Evans for priceless guidance, encouragement, patience and supporting this work and Dr. Alex P. Zijdenbos for co-supervision and providing extraordinary help in every aspect of this research. In addition, I am grateful to Vivek Singh, Jason Lerch, Yasser Ad-Dab'bagh, Oliver Lyttelton and Richard Webster for prompt and thorough feedback, corrections and valuable suggestions concerning this project; Dr. Bruce Pike, and Dr. Luis Collins for MR expertise and general understanding of MRI issues. Also, I would like to thank Dale Einarson, Anthonin Reilhac, Reza Adalat and Kelvin Mok for the software support and general help in this project; Dr. Keith J. Worsley and Vero Ravard for providing the help in statistics. Finally, I would like to acknowledge the enormous computational resources provided by the McConnell Brain Imaging Center of the Montreal Neurological Institute.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Magnetic Resonance Imaging (MRI) has revolutionized the modern practice of medicine. MRI is an imaging method applied in vivo and post mortem to reveal anatomical, diagnostic and functional information using combinations of radio waves and magnetic fields. The technique is applied in medicine as well as in biological and pharmaceutical research to create nondestructive, three-dimensional, internal images of the soft tissues of the body, including the brain, spinal cord and muscle. There are numerous applications where the MRI data must be converted into quantitative measurements or brain anatomy using image processing algorithms. Both raw data and analysis algorithm introduce errors into the measurement. It is therefore essential to characterize the performance of these algorithms across a range of error sources and algorithm parameters.

The objective of this thesis is to design an automatic, generic, standard, extensible pipeline for objective and quantitative validation of MRI classification algorithms. The scope of this project will be limited by the following constraint:

- Use of only 3 discrete and fuzzy tissue classifications: gray matter, white matter, cerebro-spinal fluid.

This introductory chapter revises the basic elements of MRI acquisition and analysis. Chapter 2 introduces the general validation theory in medical imaging and reviews previous work. Chapter 3 describes the BAT validation methodology, datasets and metrics. Chapter 4 describes the BAT design and interface. Chapter 5 presents the BAT results for some practical issues of using MRI tissue classification techniques. Conclusion and description of future work are presented in Chapter 6.

1.1 Basic Principles of Magnetic Resonance Imaging

This chapter will give a brief review of MRI principles. Other classical sources should be consulted for more details [Nishimura, 1993; Sprawls, 1992; Plewes and Bishop, 1992; Allen, 1992].

The human body primarily consists of fat and water: the major hydrogen containing components that make the human body approximately 63% hydrogen atoms. The hydrogen atom, as part of water molecule, has a single proton and hence possesses nuclear momentum. These charged nuclei, which are also called spins, have small magnetic moments and can be viewed as magnetic dipoles. Under normal conditions, when there is no external magnetic field present, the hydrogen atoms are oriented randomly with zero net magnetic moment. When an external magnetic field B_0 (main magnetic field) is applied, two groups of orientations appear: aligned and unaligned with external magnetic field (Figure 1.1). The ratio between the first and second group are described by Boltzmann statistics, and in equilibrium and at normal temperature are equal to 0.999993. The excess of nuclei in the first group produces a net magnetization vector (NMV) M oriented along the external magnetic field. Moreover, in the presence of an external magnetic field, the spins precess at Larmor frequency defined as $\omega=\gamma B$, where γ, the gyro-magnetic ratio, is dependent on the type of nuclei. For hydrogen, γ = 63.9 Mhz/T at 1.5 Tesla external magnetic field.

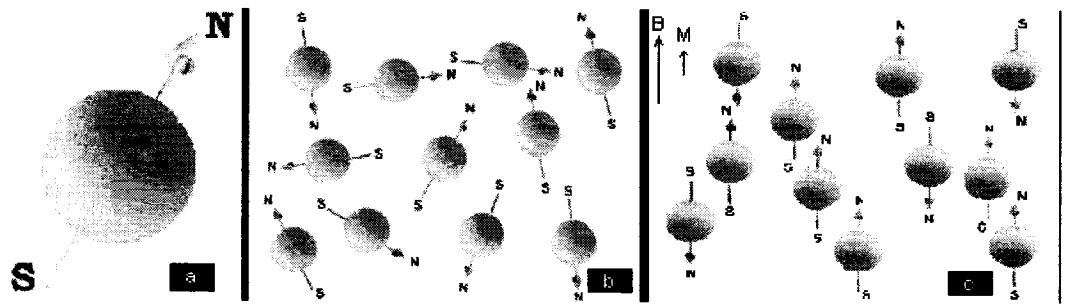

Figure 1.1 Nuclei spins and orientations: a) single spin with magnetic moment b) random orientation of spins in the absence of a magnetic field c) parallel and anti-parallel orientation of spins in an external magnetic field and net magnetization vector M in the orientation as B.

Resonance is referred to as the property of an atom to absorb energy only at the Larmor frequency. This is the basis of MR. An atom will only absorb external energy if that energy is delivered at precisely its resonant frequency. Excitation occurs when the proton absorbs the applied energy or resonates. This energy is delivered by a radio-frequency (RF) impulse, or B_1 excitation magnetic field, perpendicular to the main magnetic field. As resonance occurs, the NMV moves out of alignment with B_0 to a pre-

specified angle. The deflection of the magnetization or total angle created after the end of the RF pulse is referred to as the flip angle. The stronger the RF energy applied, the greater the angle of deflection for the magnetization. The two most common flip angles in MR are 90° and 180°. A 90° pulse will flip the magnetization into the x-y plane (Mxy). A 180° pulse will flip the magnetization through the x-y plane and into the opposite direction of B_0.

At the termination of the RF impulse, the freely precessing protons in the transverse plane (Mxy) give up energy (RF) at the same frequency that it was absorbed, in order to try to realign with B_0. As the transverse magnetization starts to decay due to the loss of phase coherence, the protons eventually realign with B_0. This signal produced by the decay (evolution) of transverse magnetization is called free induction decay (FID). The amplitude of the FID signal becomes smaller over time as net magnetization returns to equilibrium. Simultaneously, the longitudinal magnetization (Mz) begins to recover and return to an equilibrium position along B_0. Measuring this signal, during relaxation, at each space location and reconstructing the data using Fourier transform into an image is the basic MRI principle.

The relaxation properties in MR scanning are controlled by the biological parameters: spin-lattice relaxation time T_1, spin-spin relaxation time T_2 and proton densities PD. These parameters are tissue dependent, introducing the possibility to separate different tissue types in the human body. Figure 1.2 demonstrates excitation and evolution of the NMV. Figure 1.3 demonstrates an example pulse sequence and the formation of the longitudinal and transverse relaxations that produce T1, T2 and PD images.

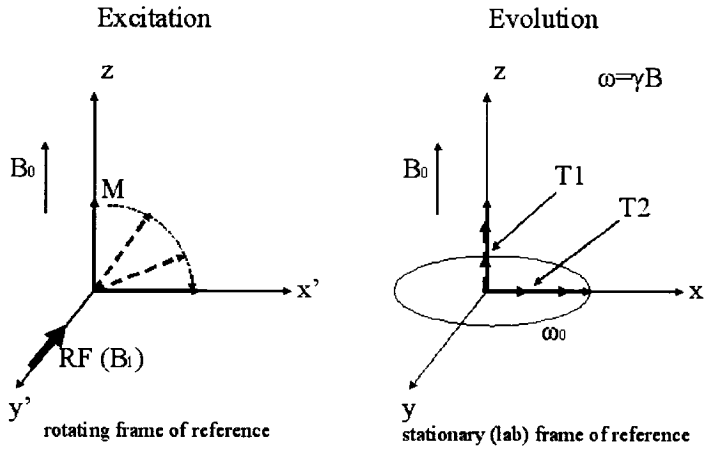

Figure 1.2 Diagram of the longitudinal and transverse components of M: Excitation - a rotating magnetic field (RF), perpendicular to Bo, with frequency ω_0 can rotate M into the x-y plane; Evolution - M will then precess freely and decay back to its equilibrium position along the z-axis. The rotation frame of the reference rotates around the z axis at the Larmor frequency ω to compensate for the spin precession and facilitate visualization and calculation.

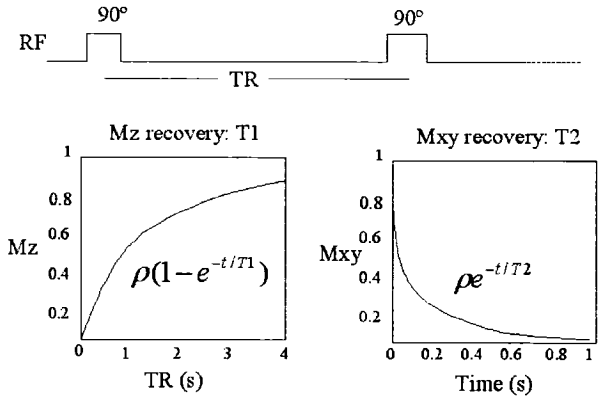

Figure 1.3 Example pulse sequence and corresponding Mz recovery time T1 and Mxy recovery time T2. These components form T1, T2 and Proton Density (PD or ρ) images. TR is the repetition time of the RF excitation pulse which flips the NMV by 90°.

Each spectrum gives different tissues contrast; for example T1-weighted image has good tissue contrast and the cerebro-spinal fluid (CSF) fluid in the ventricles and sulci appears dark. It is used for anatomical information, providing also high sensitivity for paramagnetic contrast media, fat, and fluids with high protein. By contrast, the T2-weighted image has less tissue contrast and the fluid appears bright. In brief terms, T1-weighted images give exquisite anatomical detail while T2-weighted images are generally sensitive to tissue abnormalities. Images are displayed on a grey scale format. Figure 1.4 shows T1-weighted, T2-weighted and PD MR images.

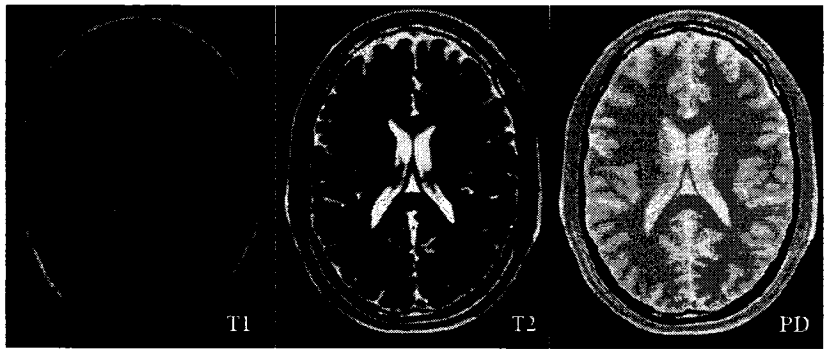

Figure 1.4 Transverse plane of T1-, T2- and PD-weighted images. The T1-weighted image has good tissue contrast and the CSF in the ventricles and sulci appears dark. The T2- and PD-weighted images have less tissue contrast and the CSF appears bright.

1.2 Sources of MR image degradation

An image artifact is any image attribute which is not present or not desired in the original imaged object. Artifacts in MRI are typically classified as to their source, such as:

- Physiological: motion, flow of blood or other fluids in the body, partial volume.
- Hardware: noise, RF and B_0 inhomogeneity, abnormal gradients, wrap around, electromagnetic spikes, receiver bandwidth limitation, sampling, averaging, voxel size.
- Inherent physics: chemical shift, susceptibility, metal.

The most important sources of MR image degradation which cause the greatest effect, from the above artifacts, on tissue classification are image noise, radio-frequency inhomogeneity, and the partial volume effect:

(i) MR image noise comes from thermal fluctuation of electronic circuits of the imaging hardware. Also, thermal variation in tissue causes generation of random RF energy contributing to the signal, producing variation in voxels intensities.

(ii) Radio-frequency inhomogeneity is due to imperfection of the RF coil. RF impulse might be not uniform across the entire field of view of the imaging object. Also, RF non-

uniformity might be caused by induced fields in imaged object. The result is intensity variation of similar tissue across field of view, so-called intensity or RF inhomogeneity.

(iii) The partial volume effect occurs due to the finite size of voxel and therefore finite resolution when one voxel represents mixed types of tissues. Most popular clinical MRI sequences produce voxel size of 1mm x 1mm x 1mm. However, transitions of tissue types at the microscopic level are smooth and gradual, rather than discrete, at 1mm voxel size level.

Simulated examples of noise and RF inhomogeneity are shown in Figure 1.5.

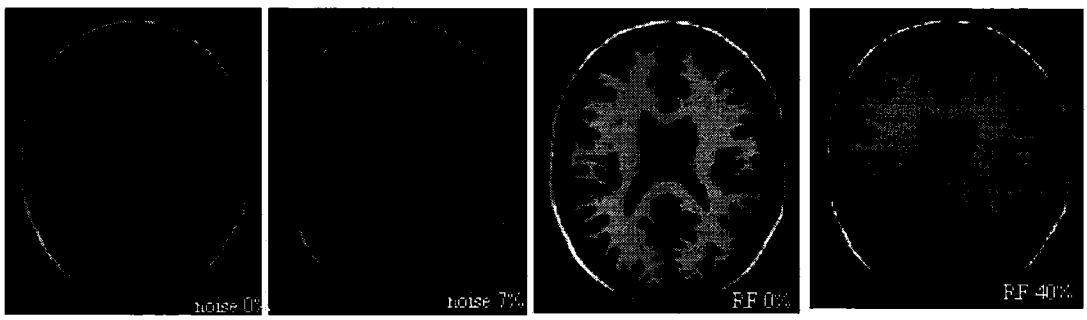

Figure 1.5 Simulated volumes with noise equal to 0%, 7% and RF inhomogeneity equal to 0%, 40%. Note that the volume with 40% RF inhomogeneity is brighter in the upper part compared to the lower. [BrainWeb, Cocosco et al. 1997].

1.3 MRI Brain Tissue Classification

Numerous tissue classification methods and their aspects are described in the literature and particularly in the project on performance analysis of automatic techniques for tissue classification by V. Kollokian [Kollokian, 1996]. Only the essential elements will be restated here.

1.3.1 Main Human Brain Tissue Types of Interest

A magnetic resonance image is a three dimensional (3D) volumetric data set, consisting of two dimensional (2D) slices where underlying anatomical information is represented by the image intensity. Each 2D slice is an array of single points or picture elements (pixels); in turn, each 3D volume is a set of volume pixels (voxels). The term "tissue

classification", in this paper, is the process of assigning each voxel in the brain image volume to one of the three major tissue types: gray matter (GM), white matter (WM), and cerebro-spinal fluid (CSF) (Figure 1.6). Although CSF is not a brain tissue, it is considered a constituent of the brain, and is presumed to be a tissue type in MRI. Other non-brain tissue types such as skin, fat, muscle, skull and other connective tissues are generally ignored.

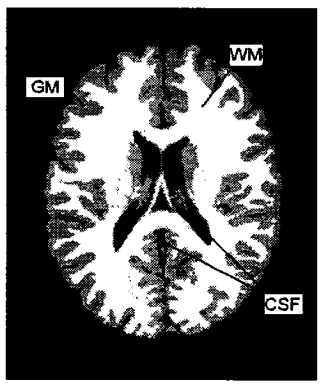

Figure 1.6 One 2D slice of a three tissue classified MRI volume: Gray Matter (GM), White Matter (WM) and Cerebro-Spinal Fluid (CSF).

1.3.2 Common Classification Techniques

There are large numbers of MRI classification techniques for normal and pathological cases. Common classification methods can be broken down in two broad subdivisions: supervised and unsupervised classifications. In a supervised classification, the analyst predicts output attributes given the values of the input attributes or training data. Unsupervised classification does not require any prior information and is based only on numerical information of the input volume. Figure 1.7 demonstrates the different type of classification techniques.

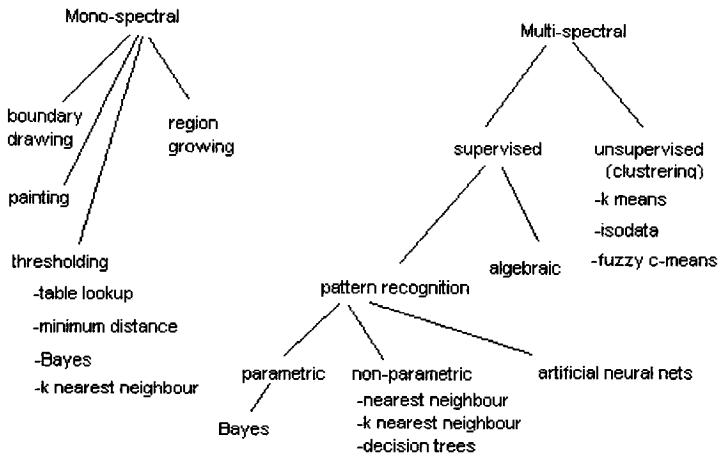

Figure 1.7 Common classification methods (Collins L. course notes on classification)

In multi-spectral imaging systems like MRI, a certain number of features representing the underlying anatomy can be used to construct 2D or 3D multi-dimensional histograms (scatter plot), where each axis represents an intensity feature. These features can be supplied by the multispectral data (e.g. T1-,T2- and PD-weighted images), where several different gray scale images of the same anatomy are obtained by different pulse sequences which yield different frequency images at each spatial location [Vannier et al., 1985; Vannier et al., 1987; Vannier et al., 1988; Vannier et al., 1991; Cline et al., 1990; Gerig et al., 1992; Hall et al., 1992; Clarke et al., 1993]. These frequencies can be grouped together in D-dimensional feature vector or multi-dimensional feature space and supplied to a classification algorithm to partition this feature space into distinct classes of interest [Schalkoff, 1992]. Clusters in the scatter plot, representing the similar tissue types, can be used to mark boundaries that allow algorithms to classify the multi-spectral data (Figure 1.8). Ideally, the clusters on the scatter plot would be well separated, but in practice, the image artifacts cause overlapping and create challenges for the classifier algorithm to make the correct decision.

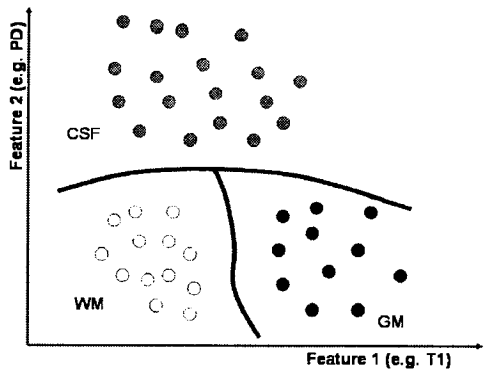

Figure 1.8 Example of cluster plot constructed from two features of MRI. Clusters in the scatter plot, representing the similar tissue types, can be used to mark boundaries that allow algorithms to classify the multi-spectral data.

An important role in supervised classification is the extraction of training data points. Training points can be manually chosen by an expert who picks the voxels or draws the regions in a brain image from an area representing a specific tissue class of interest. However, an expert tends to choose only unquestionable points, ignoring parts of the brain with uncertain regions caused by image artifacts. Therefore, a classification algorithm is provided only with the most certain points and there is no information on how to deal with ambiguous regions of the brain. Moreover, manual selection is a very tedious task with poor reproducibility. To overcome these limitations and use a full automatic classification process, the Tissue Probability Map (TPM) concept has been created that provides the prior knowledge on spatial tissue distribution [Evans et al., 1994; Kamber et al., 1995]. This concept includes the stereotaxic space (Talairach space), developed by Talairach and Tournoux [Talairach et al., 1967; Talairach and Tournoux, 1988] and tissue specific probability values denoting the level of certainty with which a particular voxel in 3D stereotaxic space belongs to one of the tissue classes, such as GM, WM or CSF. Such TPMs can be used to select training data points for fully automatic supervised classification procedures.

1.3.3 Applications of Tissue Classification Methods

MRI tissue classification is the basic step for many applications, such as the quantitative analysis of tissue volume in healthy and diseased populations [Collins et al., 2001;

Rapoport et al., 1999; Zijdenbos et al., 1998, 2002], cortical thickness measurements [Fischl and Dale, 2000; Jones et al., 2000; MacDonald et al., 2000], morphological analysis and voxel-based morphometry [Paus et al., 1999; Wright et al., 1995], as a tool for evaluation of certain diseases like Multiple Sclerosis [Kamber et al., 1992; Kamber et al., 1995], schizophrenia [McCarley et al., 1999], and Alzheimer's Disease [Tanabe et al., 1997], to evaluate effect of drug therapy on lesion or tumor load [Cline et al., 1987], build tissue probability maps [Evans et al., 1994; Kamber et al., 1995], generate phantoms for MRI and Positron Emission Tomography (PET) simulation studies [Ma et al., 1993; Kwan et al., 1996] and visualization. For all these applications, it is essential that the tissue classification method provides accurate, reliable, robust and reproducible results. However, to find the optimal classification method a thorough evaluation and validation is required.

1.4 Importance of Objective Validation of MRI Classification Algorithms

There are many different types of MRI tissue classification algorithms, based on different methodology and techniques. Over the years, many variations of these algorithms have been developed in order to improve or adapt the methods for particular needs. Given the variety of MR image classification processing methods, choosing an appropriate algorithm for an existing or a new problem can be quite a challenging task. Therefore, a method of objective validation is necessary to provide the intrinsic characteristics of the methods, evaluate their performances and limitations. Moreover, while developing a new method, validation is essential in order to compare new and existing methods and estimate the optimal processing parameters. However, since MR imaging, like many other medical modalities, is an in vivo method, validation becomes even more challenging and its complicating issues are often overlooked. For example, Prechelt [Prechelt, 1996] in his study of nearly 200 articles on neural network learning algorithms, published in 1993 and 1994 in well-known journals, observes a general lack of comparison with other algorithms, noting that most of them have serious experimental

deficiencies. His survey found that a high percentage of the new algorithms (29%) were not evaluated on any real problem at all, and that very few (8%) were compared to more than one set of alternative real data. One third of them did not present any quantitative comparison with previously known algorithms at all. Furthermore, Buvat and others [Buvat et al., 1999] emphasized that method evaluation is not enough; validation must be performed according to a specified evaluation protocol. Without all the requirements of objective validation such as standardization, statistical foundation, quantitative evaluation, validation data set and metrics, it remains difficult to compare the performance of different methods or systems and even occasionally to really understand the results of the validation process [Jannin et al., 2002; Yoo et al., 2000; Bowyer et al., 2001; Salzberg, 1997].

2. Review of Validation Theory

According to the definition by www.answers.com a validation is "a process of determining the degree to which a model is an accurate representation of the real world from the perspectives of the intended uses of the model".

2.1 Validation Theory in Medical Imaging

Significant progress in validation methodology has been made in medical image processing and Image Guided Therapy (IGT) [Goodman 1998; Shtern et al., 1999; Cleary et al., 1999; Shahidi et al., 2001; Jannin et al., 2002]. This experience can be adapted to develop an assessment methodology of MRI tissue classification methods and pipelines. Fryback and Thornbury [Fryback et al., 1991] proposed a six levels hierarchical model to appraise the efficacy of diagnostic imaging: technical capacity, diagnostic accuracy, diagnostic impact, therapeutic impact, patient outcome, societal impact. This thesis will focus primarily on the first two validation levels; the other levels apply to IGT systems and are beyond the scope of an engineering approach.

Basic validation requirements can be drawn from Quality Function Deployment (QFD), a management tool designed to capture the needs and priorities of the primary users of segmentation and registration algorithms [Yoo et al., 2000]. These users answered a series of questions such as, "What do you like about validation software as exists today?" and "If cost were not an issue, what capability would you ask for and why would you want it?". Figure 2.1 shows the results of the QFD analysis ranked from highest to lower priority.

Category	Score
1. Software Issue	144
2. Consensus Acceptability	123
3. Statistical Foundation	120
4. Ground Truth	110
5. Quantitative Evaluation	107
6. Robustness	101
7. Extensible Databases / data quality	68
8. Registration	65
9. Automation	61
10. Efficiency	57
11. Application	43
12. Multimodality	36
13. Resolution	26

Figure 2.1 Results from QFD voting process, Insight Subcommittee meeting on Validation. [Yoo et al., 2000]

The results in Figure 2.1 might be a suitable guideline for basic components, constraints and minimum set of requirements for designing MRI classification validation testbed. Similarly, many researchers [Shtern et al., 1999; Cleary et al., 1999; Shahidi et al., 2001; Buvat et al., 1999; Bowyer et al. 2001; Jannin et al., 2002] specify more generic categories of requirements concerning validation such as:

1) Standardization of validation methodology

2) Design of validation data sets and definition of corresponding ground truth

3) Design of validation metrics based on statistical foundation

2.2 Literature Review

Previous work on validation was concentrated on two different aspects. The first aspect covered only technical aspects of MRI tissue classification validation, such as the type of validation data and validation metrics. The second aspect described standardized

validation methodologies for medical imaging and appropriate statistical analysis for objective validation.

Numerous validation techniques have been used by several researches, each having its own advantages and disadvantages. Most of the validation methods, given below, are summarized in V. Kollokian's thesis [Kollokian, 1996] using existing papers describing various techniques of validation found in the MR imaging literature [Zijdenbos and Dawant. 1994; Clarke et al., 1995]. These works described the common quantitative validation methods including (i) physical phantoms, (ii) manual labeling, (iii) gross anatomy and histo-pathology, (iv) test sets, and (v) MRI simulation.

(i) Similarly to synthetic images in classical image processing assessment, physical phantoms can be used in validation of MRI classification methods. Physical phantoms are imitations of human brains and represent the cylindrical structures with compartments of known volumes, sometimes roughly shaped with different paramagnetic substances to imitate various tissue relaxation parameters [Cline et al., 1991; Gerig et al., 1992; Jackson et al., 1993; Mitchell et al., 1994]. However, it is impossible to imitate the complex spatial tissue distribution with high geometry complexity, multiple class distribution, and fuzzy volume effects of the real brain. Furthermore, physical phantoms, when placed in the MRI scanner, affect the main magnetic field differently than real human subjects and produce different RF non-uniformity. The physical phantoms have the best ground truth availability but the worst realism. Hence, physical phantoms have limited complexity and produce low realism images which make them poor choices for validation.

(ii) Another way to validate the results of MRI classifiers is to compare them with a manually labeled image, as a gold standard, produced by a human expert. However, this method has its own drawback. A human inter-operator variation is very high, as much as 40% in some cases [Zijdenbos et al., 1994] which makes the determination of a gold standard difficult. This method of validation is acceptable when a sufficient number of expert opinions determine the ground truth with statistically significant results. Unfortunately, this labor-intense process is difficult to implement for large amounts of the various data necessary for objective validation.

(iii) Gross anatomy and histo-pathology is another source of ground truth for comparing with the results of classifier. Some researchers [Taxt et al., 1992] have applied the histo-pathology of surgically removed tumors to validate the volume produced by a classifier on MRI acquired prior to the excision. However, this method is limited to pathological tissues that are marked for excision during the surgery and cannot confirm the shape and location of region of interest. Being highly labor-intensive and difficult for post-mortem analysis such as feasibility and proper excision, this method is highly impractical especially for normal brain tissues.

(iv) Test set (cross-validation, holdout method, k-fold cross-validation, leave-one-out validation etc.) is a traditional method of validation in image processing and pattern recognition. This method is based on the precondition that the data with a ground truth is separated in two disjoints parts: one part is a training data set and the second part is a testing data set. In brain imaging the training data and test data have to be provided manually by an expert which makes this method highly dependent on an expert, who tends to choose only typical intensity voxels and ignore partial volume voxels. Therefore, the test set is often underrepresented.

(v) Computer simulations traditionally provide a notion of ground truth in numerous research studies involving computer modeling. The MRI simulator developed at the BIC is based on the digital phantom [Collins et al., 1998] which represents a ground truth and MRI simulator software [Kwan et al., 1996]. MRI simulation allows the conduction of research on pulse sequence developing, noise and RF inhomogeneous modeling. Kollokian used this method in his thesis research to analyze the performance of automatic classification techniques [Kollokian, 1996]. Since all image parameters and artifacts, such as noise, resolution, RF inhomogeneity, modality, and lesion availability can be controlled by a simulator, this method is an excellent candidate for MRI validation. The disadvantage of this system is that some subtle non-linearity in MR images such as gradient field inhomogeneity cannot be mathematically described [Peterson et al., 1993], only approximated.

All above methods are based on the use of a ground truth against which a classified image will be compared for computation of the quantitative validation metrics.

However, the real MRI clinical datasets, with the highest realism, do not have any ground truth. Therefore, such data requires that a ground truth has to be constructed by labor-intensive, manual labeling or automatically, by computing a probabilistic estimate of the true classification from a collection of classifications results and a measure of the performance level represented by each classification [Warfield et al., 2004]. Another way is to use latent class analysis, that is a statistical method that estimates the accuracy, sensitivity and specificity (as latent variables) with or without a ground truth [Siu and Zhou, 1998]. This complex model requires at the least several tests to be applied to the same data in order to yield enough degrees of freedom to estimate all the parameters. These methods, for data with absence of ground truth, have advantages that can be applied to the various real clinical data and do not require laborious manual labeling of the real data, as in (ii) above. However, the produced ground truth is not objective due to its probabilistic nature and introduces ground truth's error in the final validation result.

Most of quantitative validation metrics, used in MR imaging classification methods, are based on measurements of similarity between a classification result and a gold standard on categorical data. They are based on a confusion matrix whose elements represent counts of overlap between classification results and a gold standard. This is a popular way to assess the agreement between two experts in psychiatry [Cohen 1960; Fleiss, 1975; Bartko and Carpenter, 1976]. The confusion tables can produce numerous quantitative metrics such as Jaccard, Tanimoto, Simple matching, Russel and Rao, Dice, Kulczynski [Lourenco et al., 2004; Zhang and Srihari, 2003] and Cohen kappa [Cohen, 1960; Bishop et al., 1975] metrics. However, only the kappa metric was recommended as the measure of similarity between a classification result and gold standard due to it chance correction nature [Fleiss, 1975; Bartko and Carpenter, 1976; Kollokian, 1996; Zijdenbos et al., 1994].

Several papers describe general validation methodology in medical imaging techniques [Jannin et al., 2002; Hripcsaka et al., 2002; Shtern et al., 1999; Emam 2002; Buvat et al., 1999; Styner et al., 2002; Udupa et al. 2002; Yoo et al., 2000; Bowyer et al., 2001, Cleary et al., 1999; Shahidi et al., 2001] and statistical issues of assessment

[Salzberg 1997; Flexer, 1996; Yoo et al., 2000; Feelders and Verkooijen, 1996; Gonen et al., 2001; Dietterich, 1998].

Very few papers specifically provide a validation methodology in medical imaging for comparison of classification methods. This lack emphasizes the need for a common methodology, an appropriate validation dataset and statistical analysis of the results. Attempts have been made to develop a common validation methodology for segmentation algorithms [Yoo et al., 2000] and to make recommendations for appropriate comparison of classifiers [Salzberg 1997]. However, no research has been performed towards rigorous, quantitative and objective validation methodology with statistical foundation, specifically designed for MRI tissue classification techniques.

Prior methods of validation of MRI classification techniques contain one or more of the following deficiencies:

- Limited data set
- Lack of statistical foundation
- No common methodology, data sets and metrics.

These limitations have to be overcome in order to create an objective, quantitative and rigorous validation procedure for MRI tissue classification methods with the standardized and widely accepted validation protocol. This work attempts to develop such a validation testbed, presented through the graphical user interface (GUI), based on the combination of validation techniques specifically developed for MRI tissue classification, general validation methodology for medical imaging, and comprehensive statistical tools.

2.3 Concluding Remarks

This chapter has reviewed the validation of the MR image classification literature. It has introduced the theoretical aspect and limitations of the previous work. The next chapter will describe the BAT validation methodology, datasets and metrics.

3. Methods

The Brain Analysis Testbed (BAT) consists of three main parts: validation methodology, validation data set and validation metrics. While the validation methodology and metrics should not change over time, new validation data can be added to increase evaluation objectivity, and to reflect the real world in all practical aspects, such as acquisition, age, and natural anatomical, and morphological variability of the human brain.

3.1 Validation Methodology

In order to be objective and rigorous, a validation must be performed according to a specially specified framework or methodology. In BAT, a reference-based methodology [West et al., 1997; Jannin 2002; Jannin 2003] is used that consists of performance evaluation and analysis of evaluation results (Figure 3.1). First, the validation objective or hypothesis is defined depending upon the researcher's needs. For example, one can be interested to find out which classification pipeline is more robust with regard to noise variation or has higher accuracy. According to this validation objective, an appropriate validation data set is chosen and fed into the classification pipeline, represented as a black box. For BAT all analysis is carried out within the stereotaxic space used at the MNI (Talairach space) [Talairach et al., 1967; Talairach and Tournoux, 1988] to define a standard anatomically-based frame of reference, where brains of different sizes and shapes can be directly compared after removal of size and orientation differences. The Gold Standard (GS) is pre-computed from ground truth, a manually classified volume or simulation phantom, by linear registration to Talairach space and masking procedure to keep only the tissues of interest (see "Validation Data Set" section 3.2). The classified volume is linearly transformed into the same stereotaxic space with the GS (co-registered) and relabeled according to the tissue labels of the GS for appropriate comparison. The results of the comparison are the "primary validation metrics": kappa,

accuracy, sensitivity, specificity, volumetry and partial volume effect metric. Then, these primary metrics are used to produce "secondary validation metrics": robustness and precision. This process of performance evaluation is repeated for each tested classifier or processing parameter. Next, the chosen primary and secondary metrics (i.e. kappa and robustness) of interest are statistically compared to draw the validation result.

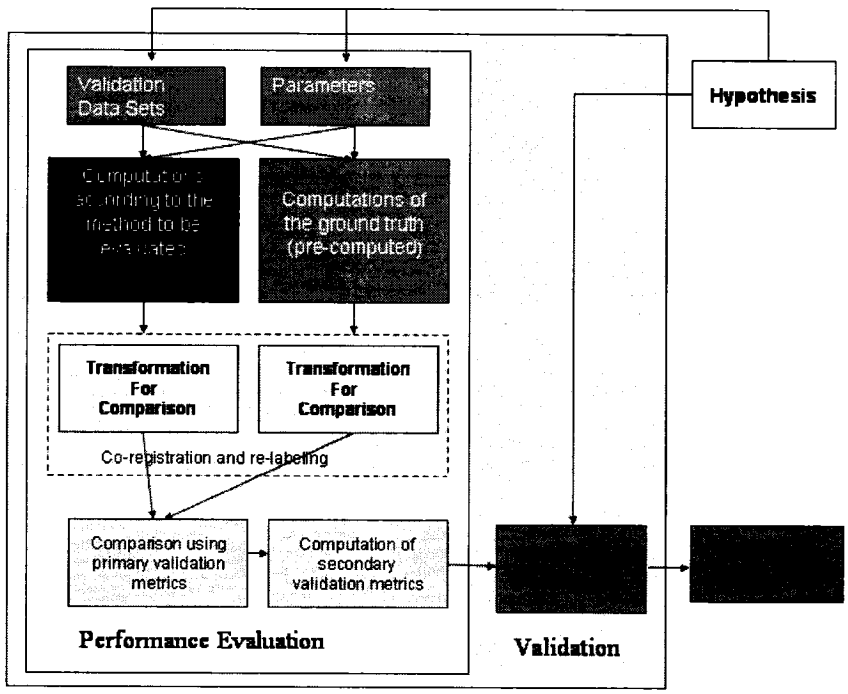

Figure 3.1. BAT design based on reference-based methodology [Jannin 2003]. The validation dataset and classification parameters are set according to the validation hypothesis. The raw image data are processed by a classification pipeline (black box). The classified volume and corresponding pre-computed ground truth are transformed into the same space, and their voxels are re-labeled using the same tissue type-label mapping for both volumes. Obtained gold standard and classified volumes are compared to produce the primary and secondary validation metrics. The process of performance evaluation is repeated for each tested classification pipeline. The validation result is attained by statistical analysis of the validation metrics for different tested methods (i.e. classifier A has a statistically higher degree of similarity with a gold standard than classifier B).

3.1.1 Statistical Issues on Comparison of Classifiers

Statistics offers many tests for measuring the significant difference between the two treatments. These tests can be adapted for comparison of two or more classifiers but

this adaptation must be done carefully, keeping in mind that these statistical tests were not specifically designed for computational methods [Salzberg 1997].

(i) Confidence intervals estimate the range of values likely to include an estimated parameter value. For sample size N>>100, the sampling distribution is assumed to be approximately normal and confidence intervals were represented by $\bar{x} \pm z_{\alpha/2}\hat{\sigma}_x$. With a probability of 95%, the true value $\overline{X}$ of the observed mean $\bar{x}$ will be within $\bar{x} \pm 1.96\hat{\sigma}_x$, where the standard error $\hat{\sigma}_x = S/\sqrt{N}$ is estimated from the sample standard deviation S and the sample size N. If the sample size N is small, for example N<<100, it is no longer justified to assume the normality of the distribution of performance measurements and t-distribution has to be employed. Confidence intervals are stricter than the statistical test of comparison of two means: if two confidence intervals do not overlap, a comparable statistical test would always indicate a statistically significant difference.

(ii) The binominal test measures whether the proportion of two categorical dependent variables significantly differs from a hypothesized proportion and requires a classified volume for each compared classifier. A binominal test provides information if the two classified 3D volumes are statistically different but it does not provide quantitative results or tell which classifier is better. Another, nearly identical form of binominal test is known as McNemar's test [Everitt, 1977] that, instead of using an exact computation using a binominal test, employs a chi-square distribution and is simpler to compute.

(iii) The McNemar test [McNemar, 1945; Sheskin, 2000] is an extremely simple way to test marginal homogeneity in K×K tables but, as a binominal test, requires two fully-classified 3D volumes representing two different treatments. It is important to notice that according to Dietterich [Dietterich 1998], only this test has acceptable type I error (0.05 probability of incorrectly detecting a difference when no difference exists). The basic McNemar test can be demonstrated (described in http://ourworld.compuserve.com) by a 2×2 table, summarizing agreement between two raters on a dichotomous trait in Table 3.1.

Table 3.1 Sample dichotomous confusion matrix

Gold Standard	Classification		
	+	-	Row Totals
+	a	b	a+b
-	c	d	c+d
Column Totals	a+c	b+d	N

Marginal homogeneity implies that row totals are equal to the corresponding column totals, or

$$(a + b) = (a + c) \qquad (3.1)$$
$$(c + d) = (b + d)$$

Since a and d on both sides of the equations cancel, this implies that b = c; this is the basis of the McNemar test calculated as

$$\chi^2 = \frac{(b - c)^2}{(b + c)} \qquad (3.2)$$

The value χ^2 can be viewed as a chi-squared statistic with 1 degree of freedom. Statistical significance is determined by evaluating the probability of χ^2 with reference to a table of cumulative probabilities of the chi-squared distribution or a comparable computer function. A significant result implies that marginal frequencies (or proportions) are not homogeneous.

(iv) Multiple comparisons. If more than two means of performances are compared by repeated pair-wise comparison a higher probability will be created of finding one or more "significant" differences when in fact there are none. The simplest approach to deal with this multiplicity effect is to use the Bonferroni adjustment of significance level. With K categories, there are exist K - 1 independent tests. For an "experiment-wise" alpha of 0.05, the Bonferroni method would make 0.05/(K - 1) the significance criterion for each test:

$$\alpha^* = \frac{\alpha}{K - 1} \qquad (3.3)$$

(v) Another, very substantial problem with reporting significance results is the repeated tuning of the algorithms in order to make them perform optimally on at least some datasets [Salzberg 1997]. Every adjustment should be considered as a separate experiment; if one is testing 10 different parameters then the significance level would have to be 0.005 in order to obtain a comparable single experiment level of 0.05. This problem can be overcome by using a larger testing data set, use separate tuning and testing data, or use of cross validation [Salzberg 1997; Flexer 1996]. On the other hand, the algorithms may perform differently on different data sets. In these cases it might be necessary to find a set of parameters to optimize the algorithm to some particular data set.

To summarize the statistical issues, the minimum general requirements for the proper statistical classifier comparison are:

- Confidence intervals, the McNemar test for comparison of performance.
- Different training and test sets.
- Independent data set in the case of parameter tuning.
- Bonferroni adjustment to correct for multiple comparisons.

In this study, the ranking was determined with the mean value of validation metric, and statistical difference between two classifiers was insured by the use of confidence intervals. The stricter McNemar test of significance difference can be used; however, this requires full 3D classified volume for each classification pipeline to be stored in BAT.

3.2 Validation Data Set

The validation dataset can be separated into four groups, according to availability of corresponding ground truth and data realism (Figure 3.2): physical phantom, numerical simulation, manually segmented real data, and real clinical data. Physical phantoms have limited complexity and produce low realism images, which make them poor choices for validation. Conversely, a real clinical dataset has the best realism and provides a large choice of different brain types; however, it does not have ground truth definition, which makes it inappropriate for reference-based methodology used in this research.

Numerically simulated and manually segmented real datasets provide satisfactory trade-off between realism and ground truth availability, and therefore were deemed adequate for practical use in this validation research.

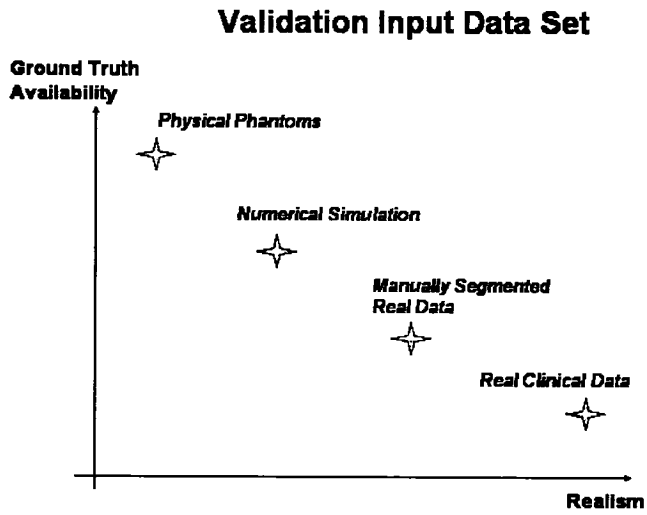

Figure 3.2 Validation dataset. There is a trade off between the realism of the data and corresponding ground truth availability.

The datasets that can be used in this project are described in the details below.

3.2.1 Physical Phantoms

The internet Brain Segmentation Repository (IBSR) (http://www.cma.mgh.harvard.edu/ibsr), provides MR images of the physical phantoms. IBSR has 3 circle shaped software phantoms (Figure 3.3) varying in signal-to-noise ratio, contrast-to-noise, and complexity.

Also, the IBSR provides the brain shaped software phantoms: 1) with noise, no gradient; 2) with noise, small gradient; 3) with noise, larger gradient; 4) with noise, large gradient; 5) no noise, small gradient. Figure 3.4 shows the original scan, phantom with gradient, and the phantom with gradient and noise.

Since the physical phantoms have limited complexity and produce low realism images, it makes them poor choice for validation; they are not used in the validation testbed.

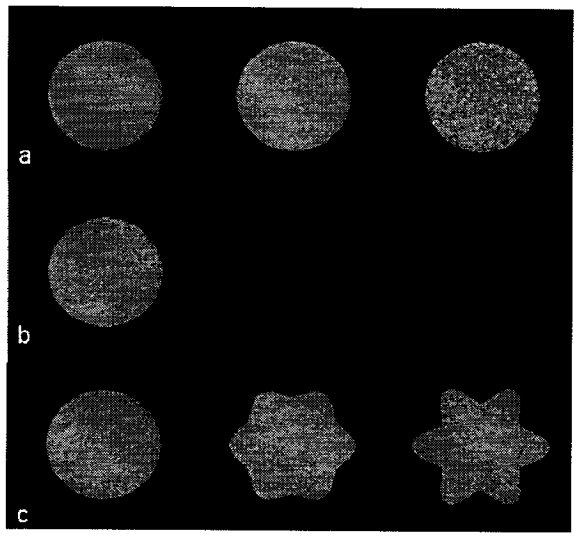

Figure 3.3 Geometric software phantoms: a) varying signal-to-noise b) varying contrast-to-noise c) varying shape complexity (IBSR, http://www.cma.mgh.harvard.edu/ibsr/)

Figure 3.4 Brain shaped software phantom (left to right): original scan, phantom with gradient, and the phantom with gradient and noise (IBSR, http://www.cma.mgh.harvard.edu/ibsr/).

3.2.2 MRI Simulated Data

An MR simulator [Kwan et al., 1996] developed at the McConnell Brain Imaging Centre allows the user to control various acquisition parameters and obtain realistic MR images of the brain. Ground truth is presented by a digital phantom that is the source for the MRI simulator. The digital phantom is a set of fuzzy volumes, in Talairach space, for the following tissue types: GM, WM, CSF, fat, muscle, skin, glial matter and connective tissue [Collins et al., 1998]. The discrete version of this digital phantom can be obtained by assigning each voxel to the most probable tissue class from the fuzzy volume. The following image parameters and artifacts were used to create MRI simulated dataset [Cocosco et. al. 1997]:

- Digital Phantom: normal adult
- MS lesions: no lesions
- Noise (%): 0, 1, 3, 5, 7, 9
- Radio Frequency (RF) inhomogeneity (%): 0, 5, 10, 20, 30, 40
- Resolution or slice thickness (mm): 1, 3, 5, 7, 9
- Modality: T1, T2, PD

The gold standard for simulated dataset was computed in the following manner:

1) GM, WM, and CSF fuzzy phantoms with 1mm isotropic voxels were discretized into one crisp digital volume.

2) Glial Matter was labeled as WM, since this project considers only 3-type tissue classification of GM, WM, and CSF.

3) All other tissue except GM, WM, and CSF were labeled as background.

4) Cerebellum was masked out with the mask constructed by the cortical surface extraction procedure [MacDonald et al., 2000], since the cerebellum structure is a mix of GM and WM.

The resulting gold standard for the young adult, shown in Figure 3.5, is a MINC (Medical Image NetCDF) format 1mm isotropic voxels size volume in Talairach space consisting only of four intensities: 0-Background, 1-CSF, 2-GM and 3-WM. Similarly, an MRI simulated dataset and ground truth is available for three types of multiple sclerosis lesions: moderate, mild and severe.

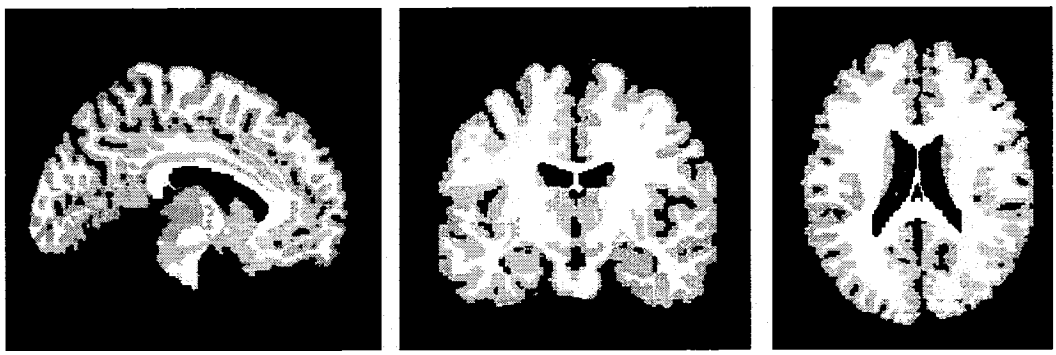

Figure 3.5 1mm Ground truth for simulated dataset consisting of GM, WM and CSF. Glial matter labeled as white matter and all other tissues except GM, WM and CSF as the background. The cerebellum was masked out because this tissue type does not belong to the three tissue classes of interest.

MRI volumes of validation input dataset with a thickness other than 1mm have to be re-sampled to 1mm to be compatible with the 1mm ground truth. The MRI simulated data was used to study the impact of different parameters of MR imaging on behavior of tissue classification techniques. Each parameter was systematically varied while keeping all others at normal, typical level: noise 3%, RF inhomogeneity 20%, slice thickness 1mm [Kollokian, 1996]. Figure 3.6 demonstrates simulated images of a normal adult with different levels of noise, RF inhomogeneity, and modality.

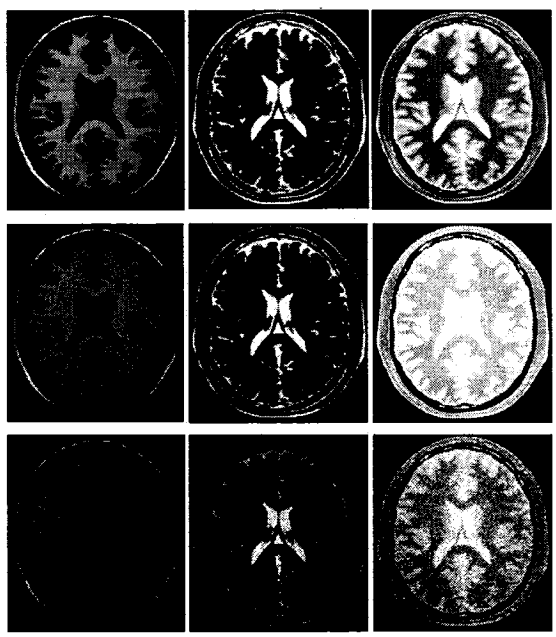

Figure 3.6. Simulated 1mm T1, T2, PD images (left to right) from BrainWeb data set [Cocosco et al. 1997]. Top: noise 0%, RF 0%; Middle (typical): noise 3%, RF 20%; Bottom: noise 9%, RF 40%.

The disadvantage of this data is that some subtle non-linearity in the MR images, such as gradient field inhomogeneity, are not properly modeled. Furthermore, phantoms idealized tissue type distribution makes the simulated images not completely realistic (Figure 3.7).

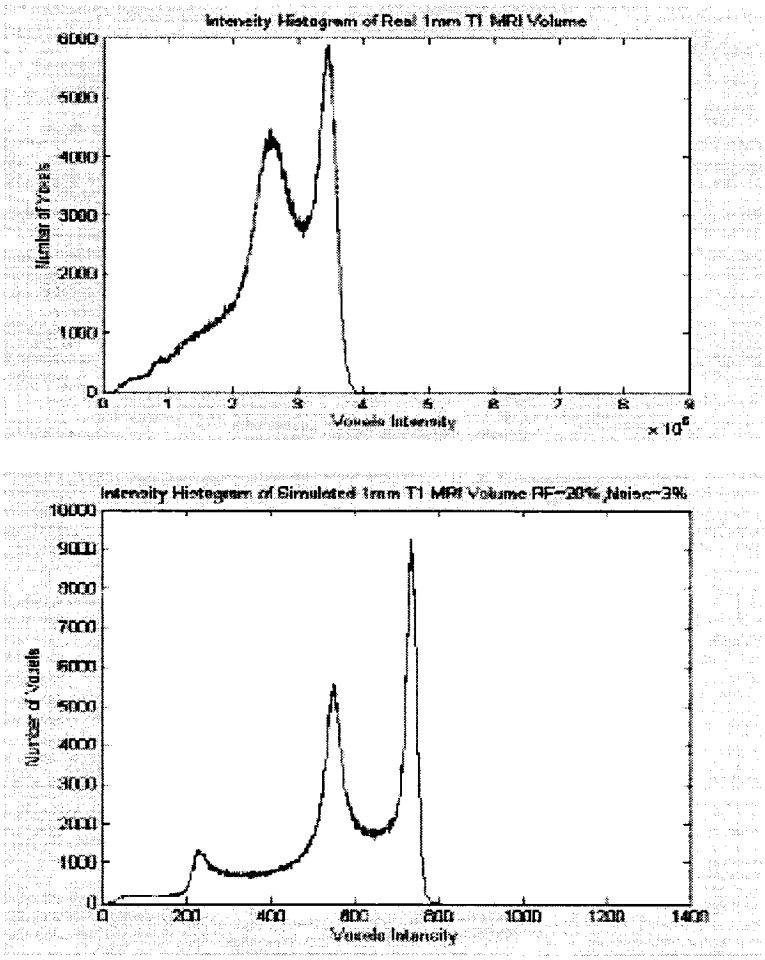

Figure 3.7 Intensity histograms of real (top) and simulated (bottom) MRI volumes. It is easy to recognize that the intensity distributions are different and much sharper for a simulated volume.

Despite these disadvantages, the simulated dataset is the main part of the validation dataset because it provides great flexibility with image parameters, as well as artifacts, and has the adequate realism-ground truth relationship. Eventually, these simulations provide necessary, but not sufficient conditions for declaring the classification method "valid", thus, real data is still needed.

3.2.3 Real Data with Ground Truth

This type of validation data is real MRI data with manually classified volumes as ground truth. A multi-spectral 1mm isotropic voxel MRI scan of a 36 year old normal male was manually labeled by a human expert [Kabani et al., 1997; Kabani et al., 1998; Kabani and Evans, 2001]. Also T2 and PD scans were acquired with 2mm sagittal slice thickness. Both acquisitions were repeated a second time, with a 1mm offset; the two paired scans were co-registered and averaged together in order to improve the image resolution. The gold standard for this data set is the discrete manually classified volume consisting of GM, WM and CSF tissues.

An additional real MRI data set is available from the Internet Brain Segmentation Repository (IBSR), Massachusetts General Hospital:

- Adult Male: T1-weighted MR Image data with complete (GM/WM/CSF) expert segmentations
- 5 year old Child: T1-weighted MR Image data with complete (GM/WM/CSF) expert segmentations
- 20 Normal Subjects: T1-weighted MR Image data with GM/WM/other expert segmentations (3.1mm slice thickness)
- 2 Tumor patients: various scans over time
- 18 Scans: T1-weighted MR Image data with expert segmentations of 43 individual structures (1.5mm slice thickness)

The real MRI data set presents an additional challenge for automatic classification since it is more realistic than the simulated data set. However, the imperfection in construction of the gold standard introduces a bias in the validation result. Furthermore, this data set can not be used for evaluation of fuzzy classification since the gold standard is a discrete volume. Despite all these disadvantages, the real data has the best realism, thus additional real data with ground truth should be added in the future to the validation data set thus improving the overall objectivity of the evaluation.

3.2.4 Real Data without Ground Truth

It is possible to use the real MRI data without ground truth for precision measurements. This dataset can be constructed by repeatedly scanning the same normal subject in a short period of time to minimize natural aging and pathological brain changes. When these MRI volumes are classified, the variation in volumes of different brain tissues is due only to acquisition and the classification pipeline and can be transformed to a precision metric. Presently, the "Colin27" dataset [Holmes, 1998] is used, consisting of 18 T1-weghted, 1mm isotropic scans of the young adult, scanned on the same scanner at the MNI. Additional real MRI multi-scan data can be used in BAT to measure the precision in the future:

- A single subject was scanned two times within a 24 hour time window each at five different MR sites over a period of six weeks using GE and Phillips 1.5 T scanners [Styner et al., 2002]. T1-weighted image - SPGR, 0.9375mm x 0.9375mm x 1.5mm, axial slicing direction. T2-weghted, PD - FSE, 0.9375mm x 0.9375mm x 3.0mm, axial slicing direction.
- 20 scans of the same subject [Clark et al., 2005]. Day 1: 5 SPGRs & 5 MPRAGEs (interleaved), Day 2: 5 SPGRs & 5 MPRAGEs (interleaved).

3.3 Validation Metrics

All validation metrics used in this project must quantitatively measure the following basic intrinsic properties of classified MRI volume:

- Similarity measurements between a gold standard and classified volume (discrete and fuzzy) based on the spatial distribution information.
- Volumetry measure of GM, WM, and CSF tissues in the classified volume and/or the gold standard.

These primary measurements can be used to compute more complex or secondary validation metrics: Robustness, Precision, and Receiver Operating Characteristic (ROC) and Area Under ROC (AUC).

All metrics must be scaled to have a minimum value of zero and maximum value of one in order to be compatible and have adequate interaction and interpretation between each other. The weighting coefficients, to scale the volumetry, partial volume effect metrics, robustness and precision metrics to the middle of their dynamic range (0.5 level) increasing sensitivity, are chosen individually for each metric by processing the typical simulated MR image using the default processing pipeline. Moreover, all validation metrics should have Confidence Intervals (CI) in order to demonstrate statistical significance in comparison. The primary metrics in this testbed have α=0.05 or 100%(1-α)=95% confidence interval.

3.3.1 Kappa, Accuracy, Sensitivity, Specificity Metrics

Comparison of classified volume against the gold standard can be viewed as agreement measurements between two raters, employing the well known confusion matrix techniques [Cohen 1960; Fleiss, 1975; Bartko and Carpenter, 1976]. Dichotomous matrix (Table 3.1) considers only a one-class case (e.g. GM/not GM), where matrix indexes can be described as follows: a - number of true positive (TP) voxels, d - number of true negative (TN) voxels, b – number of false negative (FN) voxels, c – number of false positive (FP) voxels. Similarly, in case of more than one class, the dichotomous matrix can be extended to a polychotomous matrix. Table 3.2 shows a sample polychotomous confusion matrix. In this case, N number of voxels, representing the entire volume, is separated by T number of classes designated by $C_1, C_{2,\ldots,} C_T$ where each cell n_{ij} indicates the count of voxels that the gold standard labeled this voxel as the i^{th} class, while the classifier labeled this voxel as the j^{th} class. Agreement is represented by diagonal elements n_{ii}.

Table 3.2. Sample polychotomous confusion matrix.

Gold Standard	Classification				
	C_1	C_2	...	C_T	Row Totals
C_1	n_{11}	n_{12}	...	n_{1T}	n_{1+}
C_2	n_{21}	n_{22}	...	n_{2T}	n_{2+}
:	:	:	⋱	:	:
C_T	n_{T1}	n_{T2}	...	n_{TT}	n_{T+}
Column Totals	n_{+1}	n_{+2}	...	n_{+T}	$n_{++}(=N)$

Any polychotomous confusion matrix can be represented as a set of T dichotomous confusion matrices by collapsing the matrix on each class of interest. Table 3.3 shows a sample polychotomous confusion matrix collapsed on class C_1. The elements of respective dichotomous matrix can be calculated as follows:

$$a = n_{ii} \qquad b = n_{i+} - a \qquad c = n_{+i} - a \qquad d = N - (a+b+c) \tag{3.4}$$

Coefficients of agreement can be calculated for each individual class from the respective dichotomous confusion matrix as follows [Williams, 1987]:

$$Sensitivity = \frac{TP}{S} \qquad Specificity = \frac{TN}{H} \qquad Accuracy = \frac{TP+TN}{S+H} \tag{3.5}$$

where, S=TP+FN and H=FP+TN. Sensitivity is the proportion of true positives identified voxels, specificity is the proportion of true negatives identified voxels. Note that the accuracy can be represented as

$$\begin{aligned} Accuracy &= Sensitivity\left(\frac{S}{S+H}\right) + Specificity\left(\frac{H}{S+H}\right) = \\ &= \frac{TP+TN}{TP+FN+FP+TN} = \frac{TP+TN}{N} \end{aligned} \tag{3.6}$$

Table 3.3 Sample polychotomous confusion matrix collapsed on class C1

Gold Standard	Classification		
	C_1	$C_2 \ldots C_T$	Row Totals
C_1	a	b	a+b
C_2 : C_T	c	d	c+d
Column Totals	a+c	b+d	N

These metrics do not take account of the agreement between GS and classifier due to chance. Therefore, Cohen [Cohen, 1960] developed a chance-corrected percent agreement coefficient called kappa. This metric [Bartko, 1991] has been used by numerous researchers as a similarity measurement between two labeled volumes [Fleiss, 1975; Bartko and Carpenter, 1976; Zijdenbos et al., 1994; Kollokian, 1996; Cocosco, 2002]. Kappa can be computed from the same dichotomous confusion matrix as follows:

$$k = \frac{P_o - P_c}{1 - P_c} \quad where \quad P_o = \sum_{i=1}^{T} \frac{n_{ii}}{N} \quad and \quad P_c = \sum_{i=1}^{T} \frac{n_{i+} n_{+i}}{N^2} \tag{3.7}$$

P_o is a percent agreement and P_c is the proportion of agreement due to chance.

The numerator and denominator of the overall Kappa, for several classes, can be calculated by summing the respective numerators and denominators from classes of interest in the confusion matrix defined as follows [Bishop et al., 1975]:

$$\hat{k}_i = \frac{N n_{ii} - n_{i+} n_{+i}}{N n_{i+} - n_{i+} n_{+i}} \tag{3.8}$$

where $\hat{k}_i$ is the maximum likelihood estimate of the conditional agreement between observers for a given category. Therefore, kappa per class or several classes can be calculated. For example, the overall Kappa for GM (class 2) and WM (class 3) is:

$$\hat{K}_{2,3} = \frac{\sum_{i=2,3} N n_{ii} - n_{i+} n_{+i}}{\sum_{i=2,3} N n_{i+} - n_{i+} n_{+i}} \tag{3.9}$$

The Kappa asymptotic variance that is used for CI calculation can be approximated [Bishop et al., 1975]:

$$\sigma_{\infty}^{2}[\hat{K}]=\frac{1}{N}\left\{\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2}+\frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3}+\frac{(1-\theta_1)^2(\theta_4-4\theta_2^2)}{(1-\theta_2)^4}\right\} \quad (3.10)$$

$$\theta_1=\sum_i p_{ii},\theta_2=\sum_i p_{i+}p_{+i},\theta_3=\sum_i p_{ii}(p_{i+}+p_{+i}),\theta_4=\sum_{i,j} p_{ij}(p_{j+}+p_{+i})^2$$

where $\hat{k}_i$ is maximum likelihood of kappa and $p_{ij}=\frac{n_{ij}}{N}$. Since $\hat{K}$ is asymptotically normal, we can use $\sigma_{\infty}^{2}[\hat{K}]$ to construct confidence intervals of kappa metric:

$$\hat{K}\pm Z_{\alpha/2}\cdot\sigma_{\infty}[\hat{K}] \quad (3.11)$$

where $Z_{\alpha/2}$ is the α/2-level normal deviate and equal to 1.96 for α=0.05 or 100%(1-α)=95% confidence interval. Kappa equal to one indicates perfect agreement, and kappa equal to zero indicates agreement due to chance alone. In rare situations, kappa can be negative; this is a sign that the two observers agreed less than would be expected by chance.

3.3.1.1 Kappa issues

Kappa statistics are appropriate for testing whether agreement exceeds chance levels for binary and nominal ratings. However, there are some problems and issues regarding the appropriate use of kappa statistic (http://ourworld.compuserve.com):

- Kappa is not really a chance-corrected measure of agreement due to the fact that the raters are not statistically independent, which is required for a proper calculation of proportion of chance agreement.
- Kappa is an omnibus index of agreement: it does not make distinctions among various types and sources of disagreement. Kappa treats the voxels similarly, regardless of where in the image the voxels are.
- Kappa is influenced by trait prevalence (distribution) and base-rates [Hripcsak and Heitjan, 2002]. As a result, kappas are seldom comparable across studies, procedures, or populations. Kappa value can only be interpreted in a proper

fashion when both prevalence and overall agreements are mentioned in a reproducibility study report.

Prevalence and bias indexes can be calculated as:

$$\Pr evalence = \frac{(a-d)}{N}$$
$$Bias = \frac{(b-c)}{N} \tag{3.12}$$

Prevalence is the difference between probability of 'Yes' and 'No' (cells a and d in Table 3.1). Generally, when there is a large prevalence index, kappa is lower than when the prevalence index is low or zero. The effect of prevalence on kappa is greater for large values of kappa than for small values [Sim and Wright, 2005].

Bias is the difference in proportions of 'Yes' for two raters (cells b and c in Table 3.1). When there is a large bias, kappa is higher than when bias is low or absent. In contrast to prevalence, the effect of bias is greater when kappa is small than when it is large [Sim and Wright, 2005].

In the BAT, only the same type of kappa metric, representing the same tissue type, can be compared across the pipelines, processing the same dataset (e.g. valid: GM kappa of pipeline A is higher than GM kappa of pipeline B based on the processing of the simulated 1mm, 20% RF, 3% noise T1-weighted MR image; invalid: GM kappa of pipeline A is higher than WM kappa of pipeline A or B).

3.3.2 Volumetry

The volumetry metric is based on the volumetry measurements such as tissue class volume, number of tissue voxels (GM, WM, CSF) or GM/WM ratio. The ratio of number of GM to WM voxels gives a relative metric, which is independent on voxel size. Moreover, GM/WM ratio does not depend on CSF tissues that introduce the most errors after skull striping and masking out of the cerebellum.

In the case of the validation of data with a gold standard, the volumetry metric V can be calculated as follows:

$$V = \frac{1}{1 + d/w} \quad d = \sqrt{(x - gs)^2} \tag{3.13}$$

where d is Euclidian distance between the volumetry measurement in gold standard image gs and classified image x. The normalization volumes or weighting normalization coefficients w, represented by number of voxels (same as mm^3 for 1mm isotropic size voxels), are calculated from the typical simulated MRI classified volume: GM/WM=1.34, BCK=3001960, CSF = 371945, GM = 902912, WM = 674777.
Unfortunately, the volumetry measure for a single MRI volume is a single value and can not have confidence intervals; thus, it can not be statistically compared by indicating the significant difference between the two volumetric measures or metrics.

3.3.3 Partial Volume Effect (PVE)

Partial volume effect can be measured for classifiers that support continuous or fuzzy classification. PVE evaluation is possible only with simulated datasets, since they provide fuzzy ground truth; all other datasets usually only have discrete ground truth. The voxel intensity, in the fuzzy volume, represents the probability of this voxel belonging to this particular tissue type. PVE metric can be derived by computing average distance between the probabilities of fuzzy gold standard volume gs_{ti} and classified fuzzy volume x_{ti} for each i^{th} voxel and each t^{th} tissue type:

$$PVEmetric_t = \frac{1}{1 + \frac{\bar{d}_t}{w}} \quad d_{ti} = |x_{ti} - gs_{ti}| \quad \bar{d}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} d_{ti}, \tag{3.14}$$

where $\bar{d}_t$ is an average distance between probabilities, w=0.78 is the weighting coefficient, T = {GM, WM, CSF}, and N number of voxels that corresponds to the condition $d_{ti} \neq 0$. Confidence interval $CI_{\bar{d}_t}$ for distance $\bar{d}_t$, its sample standard deviation S_t and confidence interval for PVE metric for class t can be computed:

$$CI_{d_t} = \bar{d}_t \pm z_{\alpha/2}\frac{S_t}{\sqrt{N_t}} \quad S_t = \sqrt{\frac{1}{N_t - 1}\sum_{i=1}^{N_t}(d_{ti} - \bar{d}_t)^2} \quad CI_{PVEmetric_t} = \frac{1}{1 + \frac{CI_{d_t}}{w}} \tag{3.15}$$

The overall PVE metric and confidence intervals can be calculated by counting probability distance in all voxels and in all fuzzy volumes simultaneously:

$$PVEmetric_{all} = \frac{1}{1+\frac{\bar{d}_{all}}{w}} \quad CI_{PVEmetric_{all}} = \frac{1}{1+\frac{CI_{all}}{w}} \tag{3.16}$$

$$\text{where } \bar{d}_{all} = \frac{1}{\sum_{t=1}^{T} N_t} \sum_{t=1}^{T} \sum_{i=1}^{N_t} d_{ti} \quad CI_{all} = \bar{d}_{all} \pm z_{\alpha/2} \frac{S_{all}}{\sqrt{\sum_{t=1}^{T} N_t}}$$

$$S_{all} = \sqrt{\frac{1}{\sum_{t=1}^{T} N_t - 1} \sum_{t=1}^{T} \sum_{i=1}^{N_t} (d_{ti} - \bar{d}_t)^2} \tag{3.17}$$

Again as in the case for kappa metric, $z_{\alpha/2}$ is the α/2-level normal deviate and is equal to 1.96 for α=0.05 or 100%(1-α)=95% confidence interval.

3.3.4 Robustness

Robustness is the measure or extent of the ability of a system to continue to function despite the existence of input degradation, represented by the simulated dataset with variations in noise, RF and slice thickness, varying one parameter in time and keeping others at default values. As a result, the system response can be plotted as the set of primary kappa metrics for each volume. Robustness metric is based on the slopes of the system response and distance of each primary metric on the system response from its maximum value. Figure 3.8 demonstrates the subset of system response consisting of three output data points.

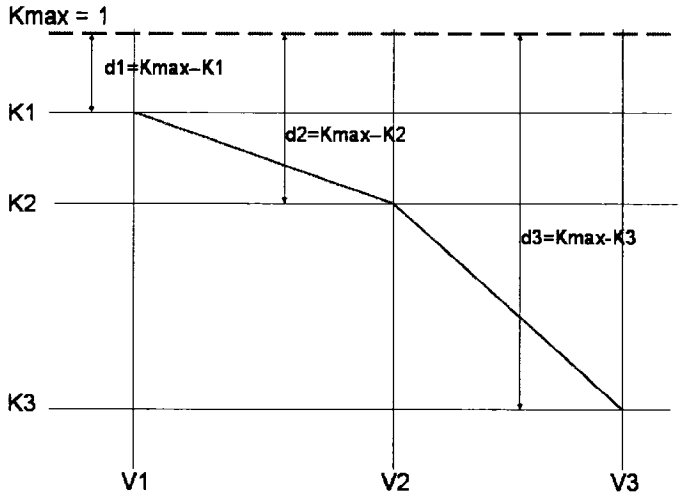

Figure 3.8 Subset of 3 output data points demonstrating the slopes of the system response and the distance from the maximum value. K1, K2, K3 are the primary kappa metrics (Kmax =1) for respective degradation of input MRI volumes V1, V2, V3. The slope characterizes the degradation of system functionality, represented by changes in kappa, due to one step in the input degradation.

Absolute value of average slope of the system response curve:

$$\bar{s} = \left| \frac{1}{N}\sum_{i=1}^{N} s_i \right| \quad s_i = \frac{\Delta k_i}{\Delta v_i} = \frac{k_i - k_{i+1}}{v_{i+1} - v_i}\ ,\ \Delta v_i > 0 \qquad (3.18)$$

where Δk_i and Δv_i are changes in the primary metrics and the input degradation between two points respectively. The average distance $\bar{d}$ between the primary metrics of the response and their maximum value is given by the equation:

$$\bar{d} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(k_i - k_{\max})^2} \qquad (3.19)$$

Using (3.18) and (3.19) robustness metric can be computed:

$$Robustness = \frac{1}{1 + \frac{\bar{s}/2 + \bar{d}}{w}} \qquad (3.20)$$

where w=0.132 is the weighting normalized coefficient. Value $\bar{s}$ is decreased by half to balance it with higher value $\bar{d}$.

To unify all input degradation for different robustness metrics such as noise, RF inhomogeneity and slice thickness, the amount of degradation is normalized for all degradation types. The degradation steps are represented by relative values equal to the

relationship of the amount of degradation to its maximum (9% for noise, 40% for RF and 9mm for slice thickness). For example, 1mm slice thickness becomes (1mm / 9mm) = 0.11, 3% noise become (3% / 9%) = 0.33, 20% RF becomes (20% / 40%) = 0.5. Normalization of input degradation provides possibilities to compare different types of robustness metrics such as noise, RF and thickness. Confidence interval provides statistical difference between two robustness metrics and can be computed by using maximum and minimum confidence values of corresponding primary kappa metrics.

3.3.5 Precision

Precision is a degree of mutual agreement among the series of individual measurements, often, but not necessarily, expressed by the standard deviation. By definition, in order to measure the precision, it is required to have a series of similar measurements N as an input. Then, these N MRI scans are classified in the same way producing N classification results represented by N volumetry measurements V. The standard deviation S, of these N volumetry measurements V, reflects the precision metric:

$$\mathrm{Pr}\,ecision = \frac{1}{1+\frac{S}{w}} \text{ , where } S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(V_i - \bar{V})^2} \text{ , } \bar{V} = \frac{1}{N}\sum_{i=1}^{N} V_i \qquad (3.21)$$

The value of w=0.042 is used for GM/WM ratio, which is equal to the standard deviation of the GM/WM ratio for the Colin27 dataset processed by the default processing pipeline.

Note that the precise result is not always accurate (Figure 3.9 top). Figure 3.9 (bottom) demonstrates classification results, presented by GM/WM ratio, for the Colin27 dataset classified by over 100 variation classification techniques used in MNI.

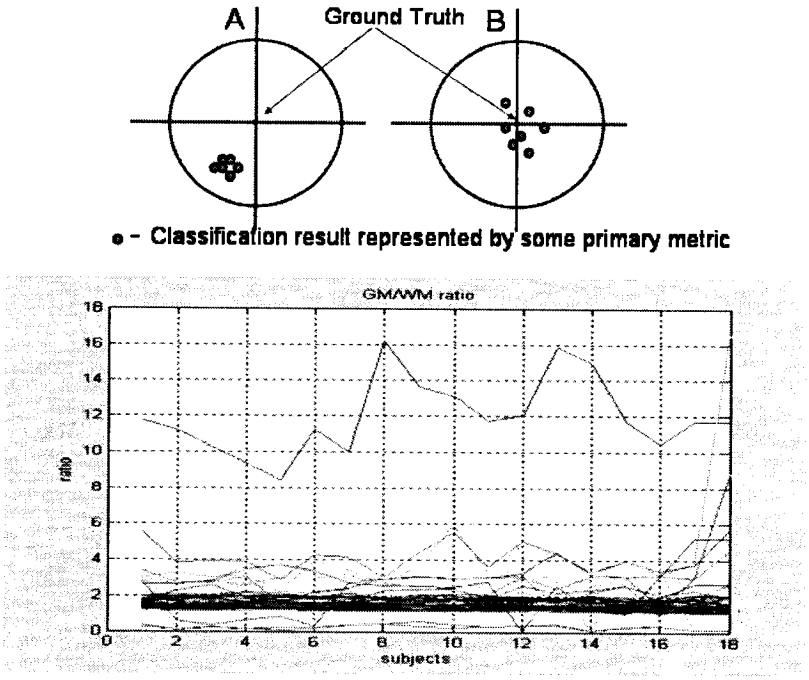

Figure 3.9 Top: A schematic illustrating the distinction between precision and accuracy among a set of independent measurements (A) Precise but not accurate – mean is far from ground truth but variability I is small (B) Accurate but not precise – mean is near ground truth but variability is large. Bottom: Illustration of variability of classification results with choice of technique, as illustrated by consistency in GM/WM ratio across repeat scans from the same brain (Colin27 database). Permutations of the classifier parameter settings and process order (BAT design section) were used to create over 100 separate classifier pipelines. Most approaches yielded a plausible ratio of ~1.4 but some yielded clearly erroneous ratios. The most precise classifier has the lowest inter-scan variance in GM/WM ratio regardless of accuracy. Therefore, the assessment of classifier quality based solely on precision could incorrectly favor an inaccurate technique.

Most of the results in Figure 3.9 (bottom) yielded a plausible ratio of 1.4, which is close to 1.26, the average value for men [Allen et al., 2002], however, some provided clearly erroneous ratios. According to the precision metric, the most precise classifier has the lowest inter-scan variance in GM/WM ratio regardless of accuracy. Therefore, the assessment of classifier quality, based solely on precision, could incorrectly favor an inaccurate technique. If the precision metric, based on the real MRI data without ground truth, and some other metrics based on the data with ground truth (i.e. kappa) were well-correlated for different types of classification techniques, then it would be possible to use the precision as a validation metric with the notion of accuracy. This would give an opportunity for objective and quantitative validation using only real MRI without the ground truth.

3.3.6 Area under Receiver Operating Characteristic

The Receiver Operator Characteristic (ROC) curve is used to assess the quality of the discriminatory power of a test using sensitivity and specificity data. This reflects the ability to distinguish between positive and negative results, and to identify the best trade-off between sensitivity and specificity. The ROC can be obtained by consistently modifying one classifier parameter or arbitrary threshold. The set of these parameters provides a set of sensitivities and specificities of the classification algorithms. In Figure 3.10, the curve A, constructed by plotting sensitivity versus 1-specificity for each output data point of system A, provides better sensitivity and specificity than systems B or C.

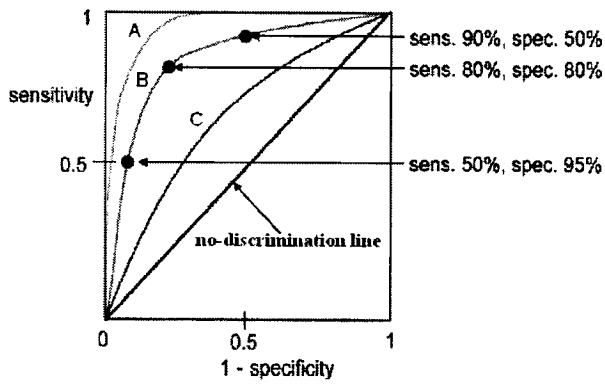

Figure 3.10 Receiving Operator Characteristic Curve examples. Classifier A has a larger area under the ROC curve than classifier B, thus classifier A outperforms classifier B. No-discrimination line represents classification by random guessing (figure from Medicopedia.com).

Each point on this curve A has higher values of sensitivity and specificity than curves B or C. The no-discrimination line represents classification by random guessing. The quantitative measure of this fact is represented by the area under ROC (AUC). Statistically speaking, the AUC of a classifier is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Since AUC is a probability, its value varies between 0 and 1, where 1 represents all positives being ranked higher than all negatives and 0.5 represents guessing. Larger AUC values indicate better classifier performance across the full range of possible thresholds. Even though it is possible that a classifier with high AUC can be outperformed by a

lower AUC classifier at some region of the ROC, in general the high AUC classifier is better.

Two methods are commonly used to compute AUC: a non-parametric method based on constructing trapezoids under the curve as an approximation of area and a parametric method using a maximum likelihood estimator to fit a smooth curve to the data points. Both methods are available as computer programs and give an estimate of area and standard error that can be used to compare different tests or the same test in different patient populations [Metz, 1978].

AUC can also be represented as the ratio of the number of correct pairwise rankings vs. the number of all possible pairs [Hand, 1997]. This quantity is called the Wilcoxon-Mann-Whitney statistic [Wilcoxon, 1945; Mann and Whitney, 1947]:

$$W = \frac{\sum_{i=0}^{p-1}\sum_{j=0}^{n-1} I(x_i, y_i)}{pn} \qquad (3.22)$$

$$I(x_i, y_i) = \begin{cases} 1 & if\ x_i > y_j \\ 0 & otherwise \end{cases}$$

where p is the number of positive points, n is the number of negative points, x_i is the i^{th} positive point and y_j is the j^{th} negative point. Generally, AUC is increased by interdependence between gold standard and test data, and diminished by noise or classification errors.

Unfortunately, ROC analysis requires a specific set of classification parameters that might be unique for each classification technique and depend on its intrinsic characteristics. Thus, there is no way to provide the uniform set of such parameters for all algorithms. This limitation makes the ROC analysis inappropriate for including in an objective and rigorous validation testbed. However, this method can serve as a visualization and optimization tool for classification performance.

3.3.7 Overall Quality Metric

Different quality metrics such as kappa, robustness and precision are produced for each brain type T, representing different populations, and averaged providing the "Global

Quality Metric for brain type T". In turn, these global metrics are averaged producing the "Overall Quality Metric" for tested classifier (Figure 3.11). The results can be stored into a database and the researcher can retrieve the best overall classifier. However, this approach has three problems and is not used presently. First of all, the validation dataset does not at the moment include all brain types equally. The second problem may arise from non-orthogonality of the validation metrics. The third problem is that ranking the classifier according to overall validation metric might be ambiguous and even meaningless, since one classifier can not provide the best performance for all data types, and no classifier is always better than any other [Wolpert, 1992].

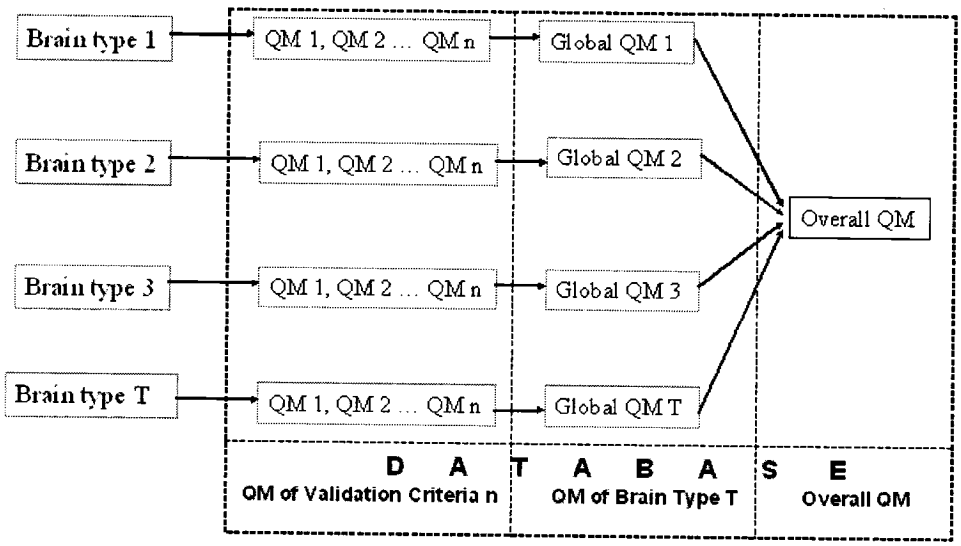

Figure 3.11 Overall Quality Metric (QM). N quality metrics (kappa, robustness, precision) are averaged producing Global QM for each brain type T. Similarly, these global QM are averaged generating Overall QM.

3.4 Concluding Remarks

The BAT validation methodology, datasets and metrics have been described.
The validation methodology is based on the reference-based approach. The validation dataset constitutes the simulated and real datasets. The validation metrics are: kappa, volumetry, pve, robustness and precision. The following chapter discusses the BAT design and interface.

4 MRI Brain Analysis Testbed (BAT) Design

Validation methodology, design of dataset, and metrics described in previous methodologies section, were implemented in the Brain Analysis Testbed (BAT) allowing the users to test the existing and new classification techniques and compare them. This testbed was created to use with the web interface (http://www.bic.mni.mcgill.ca/validation), permitting worldwide accessibility for MNI classification pipelines to produce and share the results based on the common methodology, validation dataset, and metrics.

Since, the validation datasets has to be processed or classified by the tested classification technique, two types of classification are possible depending on the location of this processing software: on-site processing and off-site processing. On-site processing is based on Montreal Neurological Institute (MNI) automated MRI processing tools (Figure 4.1) such as registration (REG), Non-Uniformity Correction (NUC), training tags extraction and cleaning, masking and partial volume estimation, supervised (Minimum Distance (MIN), Bayesian (BAYES), k Nearest-Neighbor (KNN)), and unsupervised classifications (Hard C-means (HCM), Fuzzy C-Means (FCM)).

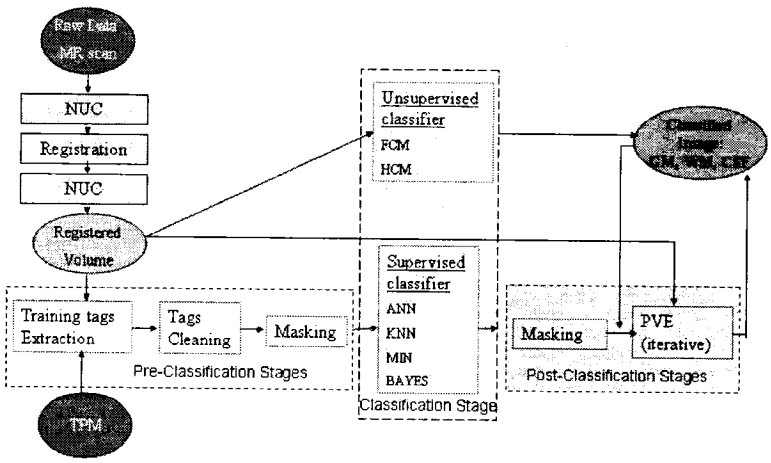

Figure 4.1. Typical MRI tissue classification pipeline configuration using at the MNI. NUC stands for non-uniformity correction stage. The Tissue Probability Map (TPM) is required for the automatic extraction of the training set used in supervised classification methods. The Partial Volume Effect (PVE) stage improves fuzzy tissue estimation and is optional. FCM, HCM, ANN, KNN, MIN, and BAYES refer to the available classification algorithms.

Off-site processing is based on the processing and classification software located on the user's site. For off-site processing, the user has to download the validation data, classify it and upload the classification result back to BAT. Then, in both cases, the processing results are validated by BAT. The validation results are stored in the BAT database, and can be displayed and compared with other classification techniques.

4.1 BAT Organization and Operation

The BAT is implemented according to the block design shown in Figure 4.2. The presentation layer is implemented on the web server and the logic layer is implemented on the BIC NFS system.

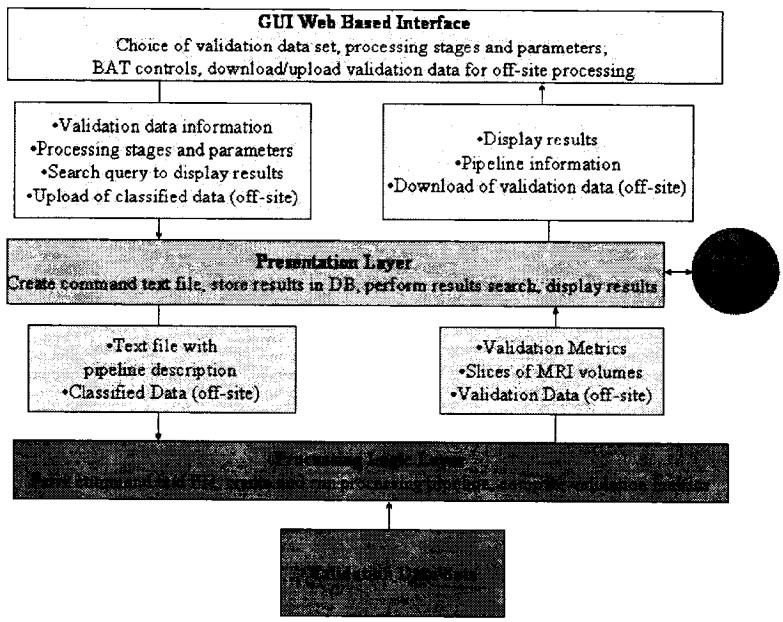

Figure 4.2. Block design of Brain Analysis Testbed (BAT): GUI web interface, presentation layer, and processing logic layer.

User web based interface allows the user to choose validation dataset of interest, process it according to the tested classification pipeline, compare, search and display the results. Information about the validation dataset and constructed processing pipeline are

coded in the text file by the presentation layer, and passes to the processing logic layer. The processing logic layer constructs the classification pipelines and selects the specified validation dataset according to this text file. When the classification pipeline is finished and validation metrics are computed, these metrics and MRI slice samples are passed back up to the presentation layer and stored in the database.

The BAT interface consists of three parts: validation dataset, processing stages, and controls. First, the user has to select the validation data that interests him or to use the preset default dataset. Then, the user has to specify on-site processing stages, or to download the validation dataset for off-site processing. On-site processing stages have stage order number, parameter field, and preset stage parameters. The value in the "stage order number" field can modify the order of the stages, or remove the stage from the processing pipeline by setting it to zero.

The parameter field accepts the valid stage parameter string to pass directly to this stage. If the parameter is not specified, then at this stage the default internal parameters are used. Comma separated parameters will be treated by a pipeline in an independent manner splitting the pipeline to execute each parameter independently. This creates the tree of pipelines where each volume will be processed separately by each branch. The format for entering the parameters is as follows: "-parameterName parameterOption". For example the set of different parameters (par1a, par1b without options; par2 with options 0.8 and 0.9) of some stage 2 in the form "-par1a,-par1b –par2 0.8,0.9" causes the splitting of the pipeline into four branches as shown in Figure 4.3.

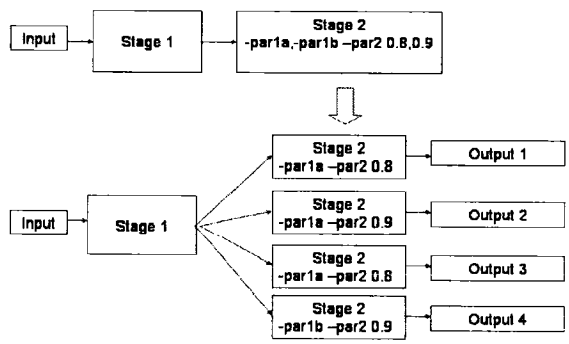

Figure 4.3 Example of stage parameters format. As the result of parameter string '-par1a,-par1b –par2 0.8,0.9' (par1a, par1b without options; par2 with options 0.8 and 0.9), stage 2 splits the input into four branches and processes the input volume separately at stage 2 with different parameters producing four outputs.

The stage parameter field should be used only by advanced users familiar with the valid set of parameters.

The custom stage allows using any other stage that exists as a script in the BIC NFS system. The stage name with a full path should be specified in the "stage name" field and the parameters of the custom stage in the "parameters" field. By clicking the "Update" button, the information from the validation dataset and processing pipeline is written to a BAT command text file. For more details on each processing stage refer to the BIC brain imaging software toolbox: http://www.bic.mni.mcgill.ca/software/. Figure 4.4 shows BAT validation dataset and processing web interface.

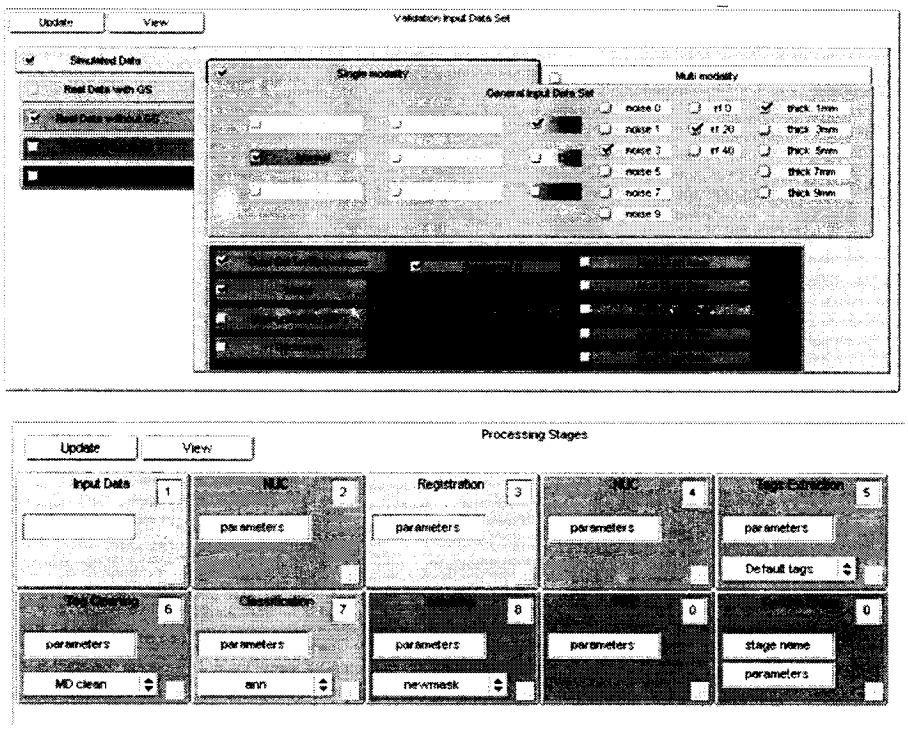

Figure 4.4. Validation input dataset form (top) and default setting of typical MNI processing stages as MRI classification pipeline (bottom).

Validation metrics are computed for each classified volume according to the provided user's tissue labels used in the classification. The final step is to specify the pipeline name and click on the "Create Pipeline" button to create a pipeline ID. Then, the user can view pipeline characteristics and decide if the created pipeline should be deleted or run.

4.2 Conclusion Remarks

This chapter has discussed the BAT design and interface. More details on the BAT organization, user instructions and updates can be found on the BAT website http://www.bic.mni.mcgill.ca/validation/. The next chapter will present some practical issues of using MRI tissue classification techniques.

5 Results Representation

When the processing pipeline is finished, the validation metrics, processing pipeline configuration, and volume slices of each processed stage for each volume are stored in the BAT database. Figure 5.1 demonstrates the visualization of one middle transverse slice of processed MRI volume for each stage.

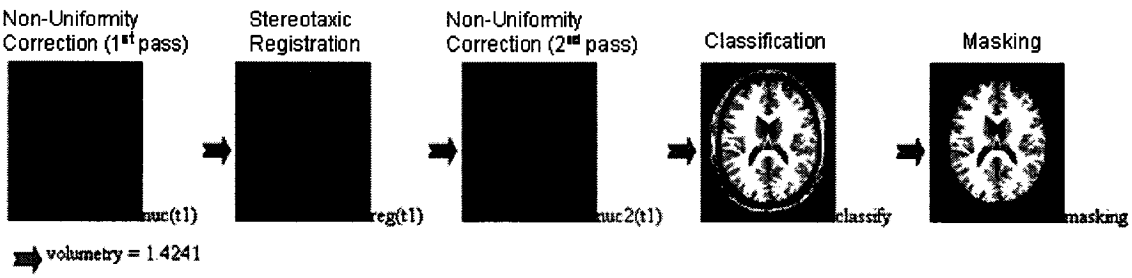

Figure 5.1. Example of progression of one T1-weighted MRI (Colin27 dataset) through the processing pipeline, displayed by the "View stages" button within BAT. The resulting metric, GM/WM volumetry in this case, is shown as the output of the validation pipeline. In this particular version of the pipeline, the non-uniformity correction (nuc) algorithm (N3, Sled et al., 1998) was run twice, once in native space and once again after the image was registered into stereotaxic space.

If processed volume has an appropriate ground truth then kappa, PVE and volumetry metrics are displayed in ascending order. If the dataset does not have a ground truth, then volumetry measure of total number of voxels of the particular tissue type or GM/WM ratio is shown for each volume. If the validation pipeline was created with robustness or a precision dataset, the secondary validation metrics obtained by this validation pipeline are presented as well. Each secondary metric is displayed together with the metric name and input volume names involved in this secondary validation metric calculation (Figure 5.2).

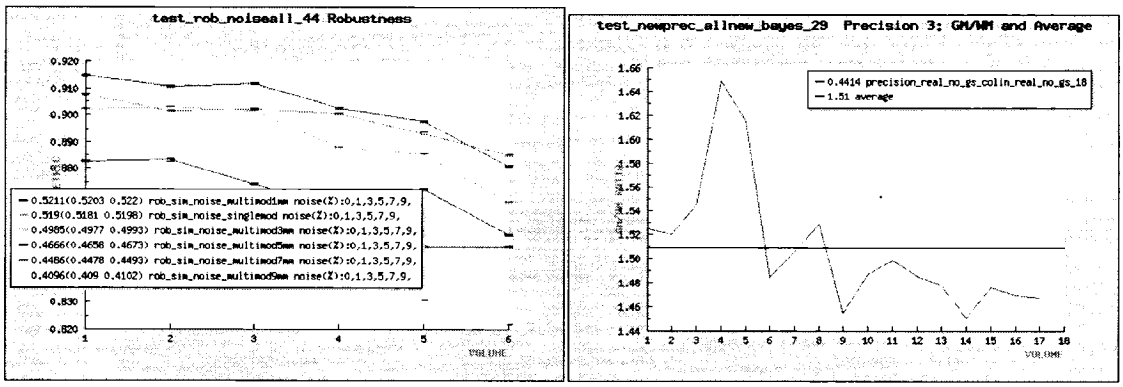

Figure 5.2. Representation of secondary validation metrics.

Left: robustness plots obtained by a sample validation pipeline. The legend shows the metric value with confidence intervals in ascending order, robustness metric name (i.e. rob_sim_noise_multimod3mm corresponds to T2/PD 3mm, T1 is always 1mm) and robustness variation parameters (here noise) and their values (here 0%, 1%, 3%, 5%, 7%, 9%).

Right: precision plot. The blue line represents validation measurements for each volume in precision dataset and the straight black line represents their average. The legend shows the metric value with corresponding precision metric name and average value.

The metric names, shown in the legend, provide information about validation dataset and degradation parameter in the case of a robustness test. Also, the user can perform detailed validation analysis and comparison based on the validation data already stored in the BAT database.

5.1 Result Search Options

The user can perform detailed validation analysis and comparison based on the validation data stored in the BAT database. This is done through the search function to display the ordered list of chosen validation metrics. The search can be performed according to validation data parameters selected in the select boxes shown in Figure 5.3.

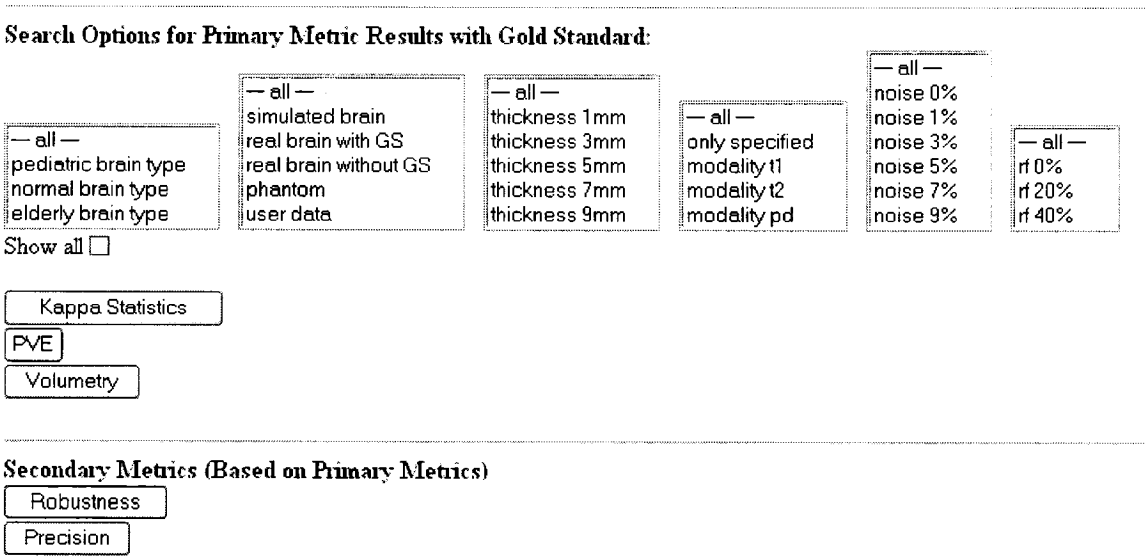

Figure 5.3 BAT search controls of validation data set parameters. After specifying desired validation data parameters, the results from the BAT database are displayed for the selected validation metric.

In the modality box, the "only specified" option indicates that only results with specified modalities should be displayed. For example, if option "modality t1" is selected only, then all results obtained by the validation data with T1 volume will be shown, even though T1 volume was a part of multimodality input together with T2 and PD. If options "modality t1" and "only specified" are selected, then only results with T1 as a single modality input will be displayed. Each box has an "all" option to select all validation data parameters described by this box. The search with "Show all", displays all validation metrics stored in the BAT database. Metrics are presented in ordered bar graph form. Each bar in this graph has an "info" link, the pipeline ID that produced this metric, and rank. The info link provides information about input volume and pipeline configuration corresponding to this metric. The significant difference between two ranks has to be verified by their confidence intervals. If the confidence intervals of compared metrics do not overlap then the metrics are significantly different with 95% probability.

When the search is performed specifying a pipeline ID then all results for this particular pipeline are highlighted by a blue color.

5.2 Practical Results Examples and Discussion

To demonstrate the validation functionality of the BAT, let us consider some practical questions regarding the use of MRI tissue classification techniques.

5.2.1 High Resolution Single-Modality versus High Resolution Mutli-Modality

It is common in MRI acquisition practice, to produce a high resolution T1-weighted image with 1mm isotropic slice thickness and low resolution T2, PD -weighted images with 3mm axial slice thickness. These simulated validation datasets, with typical values for noise of 3% and for RF of 20%, were processed by the default MNI classification pipeline using an artificial neural network classifier. The results represented as overall (GM, WM, CSF) kappa metric with confidence intervals (min, max) in the brackets:

- Single modality (T1 1mm only): kappa = 0.9021 (0.9017, 0.9024)
- Multi modality (T1 1mm, T2/PD 3mm): kappa = 0.9014 (0.9010, 0.9017)

These results demonstrate that with 95% confidence it would be better to use the single modality high resolution input rather than multi modality low resolution input data. Similarly, Figure 5.4 has been plotted by changing slice thickness of T2 and PD from 1mm to 9mm while keeping T1 at 1mm.

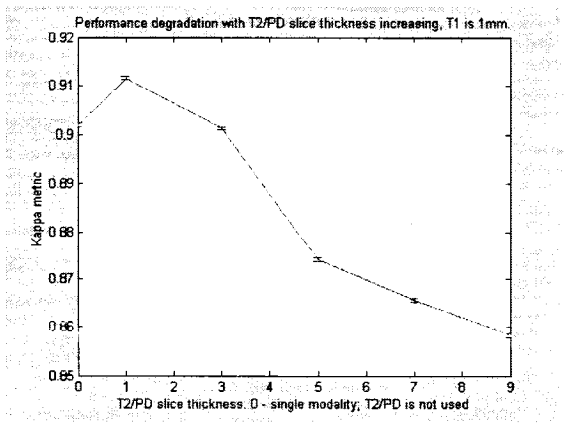

Figure 5.4. Performance degradation with T2/PD slice thickness increasing in multimodality volume with fixed T1 at 1mm. T2/PD zero thickness point corresponds to the input with T1 volume only.

The results in Figure 5.4 may suggest the best input dataset configuration in case of multimodality input dataset. For example, one can conclude that a high resolution single T1 input is preferable over a low resolution multimodality input, even though, T2 and PD volumes provide extra but nevertheless low resolution information about underlying anatomy.

5.2.2 Improving the Configuration of Processing Stages

Now let us investigate the effect of the non-uniformity correction stage (N3 algorithm) [Sled et. al., 1998] in the MNI classification pipeline. Single modality T1-weighted 1mm slice thickness volume with typical MRI artifact such as 3% noise and 20% RF inhomogeneity was classified by the default MNI processing pipeline using the ANN classification algorithm with different non-uniformity correction (NUC) stage configuration (Table 5.1).

Table 5.1. Different NUC stage modifications and corresponding kappa results (sorted by kappa). Symbol * indicates default pipeline configuration with the default NUC parameters.

Configuration	NUC parameters	Kappa
1) NUC=>REG =>NUC	-iteration 150 –stop 0.0001	0.9079 (0.9076, 0.9083)
2) NUC=>NUC=>REG	-iteration 50 –stop 0.001	0.9047 (0.9044, 0.9051)
3) NUC=>REG=>NUC*	-iteration 50 –stop 0.001	0.9037 (0.9034, 0.9041)
4) NUC => REG	-iteration 50 –stop 0.001	0.9023 (0.9019, 0.9026)

The default MNI classification pipeline has two NUC stages, one before and one after the stereotaxic registration (REG) stage (Figure 4.1, 5.1). The kappa metric shows that two NUC stages in a row (Table 5.1, line 2) provide better classification results than the default classification pipeline (Table 5.1, line 3). This result suggests that the non-uniformity correction method, used by NUC stage, does not have optimal default settings. For this input volume with 20% non-uniformity inhomogeneity, the non-uniformity correction method reaches its iteration limit before the algorithm has converged. In this instance, running the algorithm a second time before registration improves the result. It is

also possible to adjust the NUC stage parameters to change the stopping criteria and increase number of iterations by providing a new parameter string to the NUC stage in the BAT interface. Table 5.1 demonstrates that the highest kappa was obtained by using the NUC parameters "-iteration 150 –stop 0.0001". This is an illustration of configuring and finding the optimal tissue classification pipeline for some specific MRI data and processing issues. A thorough investigation of the parameter space would of course require a set of such BAT runs covering the entire range of parameter values to find the optimal setting.

5.3 Concluding Remarks

Main purposes and principles of the BAT have been demonstrated with some practical processing examples; however, the methodology presented in this research can be used to validate any other stages of an MRI processing pipeline, such as registration, masking, extraction and cleaning of the training data, partial volume estimation, and so on. The next chapter summarizes the work produced by this thesis, pointing the areas that could potentially be improved in the future.

6 Conclusion and Future Work

6.1 Conclusion

An automatic, generic, standard, extensible pipeline for objective and quantitative validation of MRI classification algorithms was designed with standardized validation protocols and associated guidelines. Validation requirements and statistical foundation for objective validation were defined.

Primary validation metrics and secondary validation metrics were designed for different degrees of evaluation. Primary validation metrics are based on the direct volumetry and similarity measurements between classified volume and ground truth. Secondary validation metrics are based on the primary validation metrics to determine a higher degree of evaluation.

Validation datasets were characterized according to the ground truth availability and data realism. Appropriate simulated and real datasets and their ground truths were selected. BAT design allows adding more validation data to reflect the real MRI data in all practical appearances and to test the MRI processing techniques in all aspects of evaluation.

Validation methodology, design of dataset and metrics were implemented in the BAT allowing the researchers to test the existing and new tissue classification techniques and to compare them. This testbed was created to be used with the web interface providing worldwide accessibility to produce and share the results based on the common methodology, validation dataset and metrics. An MRI classification was considered in the BAT as a black box, accepting two types of processing: on-site and off-site processing. On-site processing is based on Montreal Neurological Institute (MNI) automated MRI processing tools. The off-site processing is based on the user's processing and classification techniques. The tissue classification results from the black box are validated by the BAT. The validation results, stored in the BAT database, can be displayed and compared with other classification techniques.

Main purposes and principles of the BAT have been demonstrated with some practical processing examples; however, the methodology presented in this research can be used to validate any other stages of an MRI processing pipeline, such as registration, masking, extraction and cleaning of the training data, partial volume estimation, and so on.

6.2 Future Work

Several directions which may be taken to improve the existing methodology are the following:

- Introduce new validation data sets. A pediatric data set is not present in this time. It is difficult to obtain corresponding ground truth for a pediatric data set due to its high irregularity and variability of gray and white matter. The choice of old adult data is also limited by the difficulty of constructing the appropriate ground truth. Real, high resolution MRI data with ground truth obtained by post-mortem analysis will provide the best trade-off between realism and ground truth availability. An example of this real data might be the ongoing project at the Juelich Institute of Medicine Research Center. Generally, the validation data set should be as large as possible to reflect all aspects of real world MRI data and to provide a better degree of evaluation.

- Increase realism of MRI simulator. Simulated data is the best source for testing the processing methods on the data with various artifacts. However, the realism of the MRI simulator should be improved.

- Define statistical foundation for volumetry validation metrics. As described, the nature of volumetry metrics does not permit the derivation of their confidence intervals. This prevents the volumetry and precision metrics from being statistically eligible.

- Investigate the possibility of using real MRI data sets in objective validation without ground truth. Latent class analysis to estimate the accuracy, sensitivity and specificity as latent variables and an expectation-maximization algorithm for simultaneous truth and performance level estimation can be envisioned.

- Investigate and improve validation methodology. Randomization testing permits the determination and removal of optimistic statistical and methodology bias. New validation methods might reveal the hidden issues in processing pipelines and evaluate intrinsic errors associated with each processing step.

- Introduce region of interest masks to evaluate processing methods on one specific region of the brain only.

- Improve web interface to make it user friendly and facilitate the validation and comparison tasks. Continuing testing and fixing bags will improve BAT reliability.

Appendix A

Glossary and Abbreviations

acquisition The process of measuring a signal and storing it into an image file.

ANN Artificial Neural Network.

cerebellum A large structure at the lower back of the human brain. It has fine structures of intertwined gray and white matter.

cerebro-spinal fluid (CSF) Substance found surrounding the brain and within the ventricular system of the brain and spinal cord.

BAT Brain Analysis Testbed.

BIC McConnell Brain Imaging Center (MNI, McGill).

classification The process of assigning meaningful labels to different brain tissue types.

feature space A coordinate system, typically used by a classifier, where the coordinate along each axis is given by the value of a feature (a measurement).

FCM Fuzzy C-means.

ground truth A model of excellence; the reference assumed to be the "absolute true" against which a segmentation or classification result is evaluated.

gold standard Computed from the ground truth volume which is used for comparison with classified volume.

GM Gray matter: a brain tissue type that is predominantly made of neuronal dendrites.

intensity non-uniformity (INU) An artifact inherent to the MR imaging process. It is usually observed as a smooth, low spatial frequency variation in the image intensity.

KNN k Nearest-Neighbor.

kappa A chance-corrected similarity measure between two classified images.

magnetic resonance imaging (MRI) Non-invasive medical imaging technique that can produce high-resolution images with good contrast of the different biological soft tissue types.

MNI Montreal Neurological Institute (McGill University).

multi-spectral The condition denoting the nature of MR data, where underlying anatomy of the imaged organ is represented by multi-contrast images of varying characteristics.

NUC non-uniformity correction.

partial volume effect (PVE) Whenever signals from more than one tissue type are mixed in the same voxel.

phantom A digital brain model that is used both as an input to an MRI simulator, and also as a gold standard

proton density (PD) image The image acquired mostly due to proton density in the scanned region.

processing pipeline A pipeline that produce classified image from the native, original MRI image.

pipeline A set of processing stages for automatic processing of MRI images.

registration Linear spatial registration is the procedure that determines a linear (affine) transformation between two brain-based coordinate systems. If the registration is performed between an individual brain image and a standard atlas (such as Talairach's), the resulting transformation can be used to resample the individual image to the stereotaxic space defined by the atlas.

resampling The technique of changing the sampling grid of a digital image.

RF Radio-frequency pulse.

segmentation See classification.

spins The momentum of atomic nuclei with an odd number of protons and neutrons.

stereotaxic space A standard frame of reference (coordinate system) defined by anatomical landmarks of the human brain. It allows the removal of affine (translation, rotation, scale) differences between individual brains. The particular stereotaxic space used at the MNI (and in this work) is the one defined by the Talairach atlas.

supervised classification A type of classifier, where the algorithm is trained on specimen of known classes, to later classify specimen of unknown classes.

T1-weighted image (T1) An MR image in which the contrast between tissues is largely due to the longitudinal relaxation time T1.

T2-weighted image (T2) An MR image in which the contrast between tissues is largely due to transverse relaxation time T2.

tissue classification The procedure of labeling each image voxel with a tissue type GM, WM or CSF. Also called as tissue segmentation.

tissue probability map (TPM) A stereotaxic space TPM of a given tissue is a spatial probability distribution representing a certain subject population.

training tags The set of correctly labeled samples (tags) used to train a supervised classifier.

unsupervised classifier A classifier that does not required the training tags.

voxel A 3-dimensional (3D) digital image consisted of two-dimentional 2D pixel element, i.e. an image element; A 3D pixel.

validation The process of verifying if a particular classifier performed acceptable.

WM White matter: a type of brain tissue that is predominantly made of neuronal axons.

List of References

Ataman K, Street W N, Optimizing Area Under the ROC Curve using Ranking SVMs, White Paper, Department of Management Sciences The University of Iowa, (in progress).

Bartk J J, and Carpenter W T, 1976, On the methods and theory of reliability. Journal of Nervous Mental Disorders, 163(5):307-317.

Bartko JJ, 1991, Measurement and reliability: statistical thinking considerations. Schizophrenia Bulletin, 17(3):483-489.

Bishop Y M, Fiender S E, and Holland P W, 1975, Discrete Multivariate Analysis: Theory and Practice. The MIT Press, Cambridge MA.

Bowyer K W, Loew M H, Stiehl H S and Viergever M A, March 2001, Methodology of evaluation in medical image computing. Report of Dagstuhl workshop, http://www.dagstuhl.de/DATA/Reports/01111/.

Buvat I., Chameroy V., Aubry F., Pélégrini M., El Fakhri G., Huguenin C., Benali H., Todd-Pokropek A., Di Paola R., 1999, The need to develop guidelines for evaluations of medical image processing procedures. SPIE Medical Imaging, 3661, pp. 1466-1477.

Chalana V, Ng L, Rystrom L, Gee J and Haynor D, 2001, Validation of Brain Segmentation and Tissue Classification Algorithms for T1-weighted MR Images, Proc. Medical Imaging, SPIE Vol. 4322, pp. 1873-1882.

Clark K A, Woods R P, Rottenberg D A, Toga A W, Mazziotta J C, 2005, Optimizing the tissue segmentation of t1-weighted magnetic resonance images: implications for volumetric studies. Poster, HBMO 2005, Toronto.

Clarke L P, Velthunzen R, Phuphanich S, et al., 1993, MRI: stability of three supervised segmentation techniques. Magnetic Resonance Imaging, 11(1):95-106.

Clarke L, Velthuizen R, Camacho M, Heine J, Vaidyanathan M, Hall L, Thatcher R, and Silbiger M, 1995, MRI segmentation: Methods and applications. Magnetic Resonance Imaging, 13(3):343-368.

Cleary K, Anderson J, Brazaitis M, et al., 2000, "Final report of the Technical Requirements for Image-Guided Spine Procedures Workshop", April 17-20, 1999, Ellicott City, Maryland, USA. Comp Aid Surg, Volume 5, Issue 3180-215.

Cline H E, Dumoulin C L, Hart Jr H R, Lorensen W E, and Ludke S, 1987, 3D reconstruction of the brain from magnetic resonance images using a connectivity algorithm. Magnetic Resonance Imaging, 5(5): 345-352

Cline H E, Lorensen W E, Kikinis R, and Jolesz F, 1990, Three-dimensional segmentation of MR images of the head using probability and connectivity. Journal of Computer Assisted Tomography, 14(6):1037-1045

Cline H E, Lorensen W E, Souza S P, et al, 1991, 3D surface rendered MR image of the brain and its vasculature. Journal of Computed Assisted Tomography, 15(2):344-351.

Cocosco C A, 2002, Automatic generation of training data for brain tissue classification from MRI. Master thesis, Department of Electrical and Computer Engineering, McGill University.

Cohen J, 1960, A coefficient of agreement for nominal scales. Educational and Psychological measurements, 20:37-46.

Collins D L, Neelin P, Peters T M, and Evans A, 1994, Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. Journal of Computer Assisted Tomography, 18:192-205.

Collins D L, and Evans A C, 1997, ANIMAL: Validation and applications of nonlinear registration-based segmentation. Int. J.Pattern Recogn. Artific. Intell. 11: 1271-1294.

Collins D, Montagnat J, Zijdenbos A, Evans A, Arnold D, 2001, Automated estimation of brain volume in multiple sclerosis with BICCR. In: Insana, M. F., Leahy, R. M. (Eds.), Proc. of IPMI 2001. Vol. 2082 of LNCS. Springer-Verlag, pp. 141-147.

Dietterich T G, 1998, Approximate statistical tests for comparing supervised classification algorithms. Neural Computation 10, 1895-1923.

Emam K E, 2002, Benchmarking Kappa for Software Process Assessment Reliability Studies. International Software Engineering Research Network Technical Report ISERN-98-02.

Evans A, Kamber M, Collins D, and MacDonald D, 1994, An MRI-based probabilistic atlas of neuroanatomy. In Shorvon, S. et al., editors, Magnetic Resonance Scanning and Epilepsy, Plenum Press, chapter 48, pages 263-274.

Everitt B, 1997, The Analysis of Contingency Tables. Chapman and Hall, London.

Feelders A and Verkooijen W, 1996, On the statistical comparison of inductive learning methods. Artificial and Intelligence V, edited by D. Fisher and H.-J. Lenz. New York, NY: Springer-Verlag.

Fischl B, Dale A M, 2000, Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences 97 (20), 11044-11049.

Fleiss J L, 1975, Measuring agreement between two judges on the presence or absence of a trait. Biometrics, 31:651-659.

Flexer A, 1996, Statistical evaluation of neural network experiments: minimum requirements and current practice. Cybernetics and Systems '96: Proc. 13th European Meeting on Cybernetics and Systems Res., pages 1005-1008. Austrian Society for Cybernetic Studies.

Fryback D G and Thornbury J R, 1991, The efficacy of diagnostic imaging. Med. Decis. Making, 11, pp. 88-94.

Gerig G, Martin J, Kikinis R, Kubler O, Shenton M, and Jolesz F A, 1992, Unsupervised tissue type segmentation of 3D dual-echo MD head data. Image and Vision Computing, 10(6):349-360.

Gonen M, Panageas K, and Larson S M, 2001, Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. Radiology, 221:763-767.

Goodman C S, 1998, "Introduction to Health Care Technology Assessment", Nat. Library of Medicine/NICHSR, http://www.nlm.nih.gov/nichsr/ta101/ta101.pdf.

Hall L O, Bensaid A M, Clarke L P, Velthuizen R P, Silbiger M S, and Bezdek J C, 1992, A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. IEEE transactions on Neural Networks, 3(5):672-682

Hand D J, 1997, Construction and Assessment of Classification Rules. John Wiley & Sons, Bans Lane, England.

Hripcsaka G, and Heitjan D F, 2002, Measuring agreement in medical informatics reliability studies. Journal of Biomedical Informatics 35, pp. 99–110.

Jackson E F, Narayana P A, Wolinsky J S, and Doyle T J, 1993, Accuracy and reproducibility in volumetric analysis of multiple sclerosis lesions. Journal of Computer Assisted Tomography, 17(2):200-205.

Jannin P, Fitzpatrick J M, Hawkes D J, Pennec X, Shahidi R, and Vannier M W, 2002, White Paper: Validation of Medical Image processing in Image-Guided Therapy, CARS.

Jannin P, 2003, Terminology and Methodology for Validation in Medical Image Processing, MICCAI presentation, Montreal.

Jones S E, Buchbinder, B R, Aharon I, 2000. Three-dimensional mapping of cortical thickness using Laplace's equation. Hum Brain Mapp 11, 12-32.

Kabani N and Evans A, 2001, A detailed atlas of the human brain using 3D MRI. (journal tbd), (in preparation).

Kabani N, Collins L, and Evans A, 1997, Hemispheric differences in gray matter volume of adult human brain. In Society for Neuroscience Annual meeting, New Orleans-LA, USA.

Kabani N, MacDonald D, Holmes C, and Evans A, 1998, 3D atlas of the human brain. In Neuroimage (Proceedings of Human Brain Mapping 1998 Meeting), Montreal, Canada.

Kamber M, Collins D L, Shinghal R, Francis G S, and Evans A C, 1992, Model-based 3D segmentation of multiple sclerosis lesions in dual-echo MRI data. In Proceeding of the SPIE. Visualization in Biomedical Computing, volume 1808, pages 590-600, Chapel Hill, North Carolina.

Kamber M, Shinghal R, Collins D L, Francis G S, and Evans A C, 1995, Model-based 3D segmentation of multiple-sclerosis lesions in magnetic resonance brain images. IEEE Transactions in Medical Imaging, 14(3):442-453

Kollokian V, 1996, "Performance analysis of automatic techniques for tissue classification in magnetic resonance images of the human brain", Master thesis,Montreal,Concordia,McGill Universities.

Kwan R K-S, Evans A C, and Pike G B, 1996, An extensible MRI simulator for post-processing evaluation. In Hohne, K.N. and Kikinis, R., editors, Visualization in Biomedical Computing. 4th International Conference, VBC '96. Hamburg, Germany, Proceedings., volume 1131 of Lecture Notes in Computer Science, Berlin. Springer-Verlag, pages 135-140.

Lourenco F, Lobo V, and Bacao F, 2004, Binary-based similarity measures for categorical data and their application in Self-Organizing Maps. http://www.isegi.unl.pt/docentes/vlobo/Publicacoes/lobo042.pdf.

Ma Y, Kamber M, and Evans A, 1993, 3D simulation of PET brain images using segmented MRI data and positron tomography characteristic. Computerized medical imaging and graphics, 17:365-371.

MacDonald D, Kabani N, Avis D, Evans A C, September 2000, Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12 (3), 340-56.

Mann H B and Whitney D R, 1947, On a test whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statististics, 18:50–60.

Margaret M King, "Basic Principles of MRI". RT (R)(MR) http://www.erads.com/mrimod.htm

McCarley R, Wible C, Frumin M, Hirayasu Y, Levitt J, Fischer I, and Shenton M, 1999, MRI anatomy of schizophrenia. Biological psychiatry 45, pp. 1099-1119.

McNemar Q, 1947, Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157.

Metz C E. 1978, Basic principles of ROC analysis. Sem Nuc Med., 8:283-298.

Mitchell J, Karlik S J, Lee D H, and Fenster A, 1994, Classification and analysis of multiple sclerosis lesions in spin-echo MR exams. SPIE proceedings, 2359:362-372.

Paus T, Zijdenbos A, Worsley K, Collins D L, Blumenthal J, Giedd J N, Rapoport J L, Evans A C, March 1999, Structural maturation of neural pathways in children and adolescents: in vivo study. Science 283 (5409), 1908-11.

Peterson J, Christofferson J, and Golman K, 1993, MR simulation using k-space formalism. Magnetic Resonance Imaging, 11:557-568.

Prechelt L, 1996, A quantative study of experimental evaluation of neural network learning algorithms: current research practice, Neural Networks, Vol. 9.

Rapoport J L, Giedd J N, Blumenthal J, Hamburger S, Jeffries N, Fernandez T, Nicolson R, Bedwell J, Lenane M, Zijdenbos A, Paus T, Evans A, July 1999, Progressive cortical change during adolescence in childhood-onset schizophrenia. A longitudinal magnetic resonance imaging study. Arch Gen Psychiatry 56 (7), 649-54.

Salzberg S L, July 15, 1997, On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Methodological Note, Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA.

Schalkoff R, 1992, Pattern recognition-Statistical, Structural and Neural Approaches. John Wiley and Sons, New York, NY.

Shahidi R, Clarke L, Bucholz R D, et al., 2001, "White paper: Challenges and opportunities in computer-assisted interventions January 2001", Comp Aid Surg Volume 6, Issue 3, 176-181.

Sheskin D J, 2000, Handbook of parametric and nonparametric statistical procedures (second edition). Boca Raton: Chapman & Hall.

Shtern F, Winfield D et al., 1999, Report of the Joint Working Group on Image-Guided Diagnosis and Treatment. April 12-14, 1999 Washington, D.C.

Sim J and Wright C C, 2005, The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy, Volume 85, Number 3, pp.257-268.

Siu L H and Zhou X H, 1998, "Evaluating of diagnostic tests without gold standards". Statistical Methods in Medical Research, 7:354-370.

Styner M A, Charles H C, Park J, Gerig G, 2002, Multi-site ˙dation of image analysis methods – Assessing intra and inter-site variability. S₁ ıE MI-2002, San Diego.

Talairach J and Tournoux P, 1998, Co-planar stereotaxic atlas of human brain. Georg Thieme.

Tanabe J, Amend D, Schu N, DiSclafani V, Ezekiel F, Norman D, Fein G, and Weiner M, 1997, Tissue segmentation of the brain in Alzheimer disease. AJNR Am J Neuroradiol 18(1), pp. 115-123.

Taxt T, Lundervold A, Fuglaas B, Lien H, and Abeler V, 1992, Multispectral analysis of uterine corpus tumors in magnetic resonance imaging. Magnetic Resonance in Medicine, 23:55-76.

Udupa J K, LaBlanc V R, Schmidt H, Imielinska C, Saha P K, Grevera G J, Zhuge Y, Molholt P, Jin Y, Currie L M, 2002, A Methodology for Evaluating Image Segmentation Algorithms, SPIE.

Vannier M. W., Butterfield R. L., Jordan D., et al., 1985, Multispectral analysis of magnetic resonance images. Radiology, 154(1):221-224.

Warfield S K, Zou K H, Wells W M, 2004, "Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation", HBM conference.

West J, Fitzpatrick J M, Wang M Y, Dawant B M, Maurer C R, Kessler R M, Maciunas R J, Barillot C, Lemoine D, Collignon A, Maes F, Suetens P,

Vandermeulen D, Van Den Elsen P A, Napel S, Sumanaweera T S, Harkness B, Hemler P F, Hill D, Hawkes D J, Studholme C, Maintz J B A, Viergever M A, Malandain G, Pennec X, Noz M E, Maguire G Q, Pollack M, Pellizzari C A, Robb R A, Hanson D, and Woods R, 1997, Comparison and evaluation of retrospective intermodality brain image registration techniques. Journal of Computer Assisted Tomography, 21:554-566.

Wilcoxon F, 1945, Individual comparisons by ranking methods. Biometrics, 1:80–83.

Williams L E, 1987, The Diagnostic Process, volume 3. CRC Press, Boca Raton FL.

Wolpert D, 1992, On the connection between in-sample testing and generalization error. Complex systems, 6:47-94.

Wright I C, McGuire P K, Poline J-B, Travere J M, Murray R M, Frith C D, Frackowiak R S J, Friston K J, 1995, A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 2, 244-252.

Yoo T S, Ackerman M J, Vannier M, 2000, Toward a common validation methodology for segmentation and registration algorithms. Proc. Of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000), Pittsburgh, USA, Lecture Notes in Computer Science, Springer Verlag, Vol. 1935, pp. 422-431.

Zhang B and Srihari S, 2003, Binary Vector Dissimilarity Measures for Handwriting Identification. SPIE, Document Recognition and Retrieval X, Santa Clara,California.

Zijdenbos A P, and Dawant B M, 1994, Brain segmentation and white matter lesion detection in MRI images. Critical Reviews in Biomedical Engineering, 22(5&6):401-465.

Zijdenbos A P, Dawant B M, Margolin R A, and Palmer A C, 1994, "Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation", IEEE Transactions on Medical Imaging, Vol. 13, No. 4

Zijdenbos A, Forghani R, Evans A, October 1998, Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of INSECT. In: Wells, W. M., Colchester, A. C. F., Delp, S. (Eds.), Proc. of MICCAI'98. Vol. 1496 of LNCS. Springer-Verlag, pp. 439-448.

Zijdenbos A P, Forghani R, Evans A C, October 2002, Automatic 'pipeline' analysis of 3D MRI data for clinical trials: Application to multiple sclerosis. IEEE Trans Med Imaging 21 (10), 1280-91.