University of California

Los Angeles

Computer Aided Segmentation and Early Therapeutic Response Classification (CADrx) for Glioblastoma Multiforme (GBM) Brain Tumors with Magnetic Resonance Imaging

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomedical Physics

by

Jing Huo

2012

Abstract of the Dissertation

Computer Aided Segmentation and Early Therapeutic Response Classification (CADrx) for Glioblastoma Multiforme (GBM) Brain Tumors with Magnetic Resonance Imaging

by

Jing Huo

Doctor of Philosophy in Biomedical Physics

University of California, Los Angeles, 2012

Professor Matthew S. Brown, Chair

Glioblastoma multiforme (GBM) is the most common and aggressive type of primary brain tumor. Magnetic resonance (MR) imaging plays an important role in the detection of brain tumors and treatment response assessment of drugs in clinical trials. Diffusion weighted magnetic resonance imaging (DW-MRI) has the potential to work as surrogate biomarker to reveal early changes in the tumor microenvironment that precede morphologic tumor changes.

In this dissertation, we developed a computer-aided therapeutic response system (CADrx) for GBM brain tumors using T1w post-contrast MR and diffusion-weighted (DW) MR images in clinical trials. There are two components: 1) semi-automated segmentation of GBM brain tumors on T1w post-contrast MR images; 2) prediction of volumetric treatment response using early ADC values derived from the DW-MRI. The first component is the main focus of the dissertation.

The overall goal is to first facilitate radiologists in the time-consuming task of tumor contouring and generate as reproducible segmentation as possible, and then collect potential features automatically and use machine learning techniques to explore patterns in the large-scale dataset. By doing this, we aim to provide radiologists a second opinion in tumor contouring and therapy response classification.

The dissertation of Jing Huo is approved.

Whitney B. Pope

Kazunori Okada

Michael McNitt-Gray

Jeffry R. Alger

Matthew S. Brown, Committee Chair

University of California, Los Angeles
2012

To my dear Mom Wenlian Yao and Dad Xuedong Huo,

Table of Contents

List of Figures

List of Tables

Acknowledgments

The past six years at UCLA have been an unforgetful journey. I owe this to many people who contributed in many different ways to the completion of this study. Before proceeding to a long list of names, I would like to thank Professor Matthew Brown for his guidance throughout my graduate study. His enthusiasm in science and innovation, his vision in predicting the future of medical imaging, and his patience and generosity encouraged me to devote myself to the research. Without him, I could not have reached the achievements.

I am also thankful to Professor Kazunori Okada, Professor Michael McNitt-Gray, Professor Jeffry Alger, and Professor Whitney Pope for serving on my committee. Professor Okada serves as my co-advisor, and his passion about scientific research and imaging algorithm development has given me great motivation to strive for excellence. Professor McNitt-Gray, the director of our program, is taking care of each student regarding the PhD progress and the PhD training, and makes me feel I am a member of a big family. Professor Alger teaches me everything about MR imaging physics which is a difficult but fascinating image modality. Professor Pope teaches me brain anatomy and gave me the opportunity to collaborate with him showing the significance of quantitative imaging in helping clinical findings. I feel so lucky to have such a great committee and I love them!

I would like to thank my teammates: Daniel Chong, Pechin Lo, Bharath Ramakrishna, Gregory Chu for their support. The discussion and the meetings with them helped so much with my project, and without their technical supports I would not make the achievements. I would definitely give credits to the team. I also appreciate the company of friends at UCLA: Peggy, Jiayan, Zhongqi, KP, Jieying, Anna, Jinjun, He Lin, Yingkun, Xiaokui, Wei Sha. Special thanks to my best friend, Peggy, for her crazy ideas about all kinds of fun parties which makes

my life in UCLA so memorable and for her sharing all my complaints, and sharing all her funny stories.

I also would like to give my special thanks to Fusheng who encouraged me to apply for PhD program, who has taught me to try hard and not to give up when facing difficulty, who gave me a lot of inspiration on overcoming my own weaknesses and becoming a better person.

Above everyone else, I would like to share accomplishment with my parents, who always support me even when I was most bitter. To my dearest Mom and Dad, it is your unselfish love and your unconditional support that gives me the courage to go through all the difficulties. I love you!

At last, to all who have had an impact on me, whether directly or indirectly, I thank you all for what has taken place among us.

Vita

2000–2004	B.S. Electrical Engineering, Shandong University, P.R.China
2004–2006	M.S. Communications Technology, Ulm University, Germany
2006-2012	Research Assistant, Center for Computer Vision and Imaging Biomarkers, Department of Radiology, UCLA.

CHAPTER 1

Introduction

1.1 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI), or nuclear magnetic resonance imaging (NMRI), is primarily a noninvasive medical imaging technique used to visualize detailed human anatomy. MRI provides much greater contrast between the different soft tissues of the body than computed tomography (CT) does, making it especially useful in neurological (brain), musculoskeletal, cardiovascular, and oncological(cancer) imaging. The very first MR image was produced in 1973 by Nobel Prize winner Paul Lauterbur [Lau89].

MRI uses no ionizing radiation. Rather, it uses a powerful magnetic field to align the nuclear magnetization of (usually) hydrogen atoms in water in the body. Radio frequency (RF) fields are used to systematically alter the alignment of this magnetization. This causes the hydrogen nuclei to produce a rotating magnetic field detectable by the scanner. This signal can be manipulated by additional magnetic fields to obtain different responses from different tissues and construct an image of the body [WKM06].

1.1.1 Basic Principles of MRI

Generally speaking, MR images water molecules contained in the body. Each water molecule has two hydrogen nuclei or protons. Within an external magnetic field B_0, the magnetic moments of some of these protons change, and align with the direction of the field. The majority of the protons will align parallel to the field, while the minority of the protons will align anti-parallel to the field. Therefore, there will be a net effect. If we define the direction of the external B_0 as z-axis, the net magnetization vector is along z-direction. This is the signal to be manipulated and measured.

A RF wave at a certain frequency is applied to flip the net vector from z-axis to xy-plane in order to measure the signal. The "resonance" frequency is a characteristic frequency called Larmor frequency: $f = \gamma * B_0$, with γ as the gyromagnetic ratio. As the intensity and duration of the RF pulse increases, more aligned spins are affected. After the RF pulse is turned off, the protons precess and return to their equilibrium state. The vector can be decomposed into two compartments. The vector in z-axis will build up from zero to full recovery, which is called T1 relaxation, or spin-lattice relaxation. The definition of T1 is the time that it takes for the longitudinal magnetization to reach 63% of its final value, assuming a 90° RF pulse. The vector in xy-plane will decay to zero, which is called T2-decay, T2-relaxation, or spin-spin relaxation. The definition of T2 is the time that it takes for the transverse magnetization to decay to 37% of its original value. The T2-decay in the xy-plane is the signal detected by MR scanners.

A magnetic gradient field applied in three directions is used to enable spatial encoding to localize each voxel in the tomographic imaging. First, a gradient field is applied in z-axis. The Larmor frequencies at different locations are different due to different magnetic field strength. Then, phase-encoding and frequency-encoding are applied by gradient magnetic field in xy-plane. The signal will be recorded to fill in the k-space, and the inverse Fourier transform is applied to reconstruct the spatial images. Richard R. Ernst was the first to apply Fourier Transform to MR and was awarded the Nobel Prize for his contributions [EA66].

1.1.2 T1-weighted Post-contrast Volumetric MRI

Structural brain anatomy with brain tumors enhanced can be imaged by T1-weighted volumetric scans. T1 weighted image means that most of the image contrast between tissues is due to differences in the T1 value.

In order to generate MR scans, pulse sequences are programmed to obtain images with a certain weighting scheme. In a spin echo pulse sequence design, first a 90° RF pulse is applied to flip the magnetic vector into the transverse plane, and then a 180° RF pulse is applied to generate an echo signal for detection. The time between the peak of the 90° RF pulse and the peak of the echo is called the time to echo or echo time (TE), and the repetition time (TR) is defined as the time it takes to go through the pulse sequence once. In order to obtain image contrast weighted on T1 values, short TE and short TR are required. T1 weighted images can be acquired using either spin-echo or gradient-echo sequences. Gradient-echo based T1-weighted sequences can be acquired very rapidly because of their ability to use short inter-pulse TR. T1-weighted contrast can be increased with the application of an inversion recovery RF pulse.

T1 weighted MR is one of the basic types of MR contrast commonly used in clinical practice and tumor contrast is enhanced on the images. The blood-brain barrier (BBB) consists of a complex of capillary endothelial cells and serves as an effective physical barrier to the entry of toxic substances into the brain. Thus, in the normal brain tissues, contrast agent is blocked from entering the brain region. On the other hand, in highly vascularized malignant brain tumors, the BBB is disturbed and the tumor capillaries leak contrast agent into the surrounding brain tissue. The contrast agents alters the relaxation times of hydrogen protons. Contrast agents may be injected intravenously to enhance the appearance of blood vessels and tumors. The most commonly used intravenous contrast agents are based on chelates of gadolinium.

1.1.3 Diffusion Weighted MRI

Diffusion MRI is a method that produces in vivo images of biological tissues weighted with the local microstructural characteristics of water diffusion. The field of diffusion MRI can be understood in terms of two distinct classes of application: diffusion weighted MRI and diffusion tensor MRI.

In diffusion weighted imaging (DWI), the signal depends on the microscopic mobility of water. This mobility, classically called Brownian motion, is due to thermal agitation and is highly influenced by the cellular environment of water. Because water diffusion is strongly affected by molecular viscosity and membrane permeability between intra- and extracellular compartments, DW-MRI can be used to characterize highly cellular regions of tumors versus acellular regions. Treatment response detection can be manifested as a change in tumor cellularity, which may precede tumor size changes. Thus, findings on DW-MRI could be an early indicator of biologic changes. DWI is most applicable when the tissue of interest is dominated by isotropic water movement e.g. grey matter in the cerebral cortex and major brain nuclei, where the diffusion rate appears to be the same when measured along any axis.

Diffusion tensor imaging (DTI) is important when a tissue, such as the neural axons of white matter in the brain, has an internal fibrous structure leading to the anisotropy of water diffusion. Water will then diffuse more rapidly in the direction aligned with the internal structure, and more slowly as it moves perpendicular to the preferred direction. This also means that the measured rate of diffusion will differ depending on the direction of the observation. In DTI, each voxel therefore has one or more pairs of parameters: a rate of diffusion and a preferred direction of diffusion. The properties of each voxel of a single DTI image is usually represented by tensor, which is calculated from six or more different diffusion weighted acquisitions and each acquisition is obtained with a different orientation of the diffusion sensitizing gradients. In some methods, hundreds of measurements are made to construct a single image data set. The high information content of a DTI voxel makes it extremely sensitive to subtle pathology in the brain. In addition, the directional information can be exploited at a higher level of structure to select and follow neural tracts through the brain, a process called tractography.

The apparent diffusion coefficient (ADC) map, derived from diffusion weighted MR images, is the physical measurement of the water molecule movement using the following equation: $ADC = -ln[S(b) - S(0)]/b$, with b being the diffusion sensitivity factor ranging between 700 and 1000 s/mm^2, $S(0)$ and $S(b)$ being the image intensity when $b = 0$ and $b = 700 - 1000$ s/mm^2. For DWI images, three gradient-directions are applied, sufficient to estimate the trace of the diffusion tensor or "average diffusivity". For DTI, six or more gradient-directions are applied, and an tensor matrix is estimated for each voxel in the image. Based on the tensor matrix, mean diffusivity could be calculated, as well as other measurements like fractional anisotropy. Moreover, the principal direction of the diffusion tensor can be used to infer the white-matter connectivity of the brain (i.e. tractography).

1.2 Response Assessment Criteria for High-grade Gliomas

1.2.1 Glioblastoma Multiforme (GBM)

GBM is the most common and most aggressive type of the primary brain tumor. The current World Health Organization (WHO) classification of primary brain tumors lists GBM as a grade IV infiltrative glioma. GBMs are the most common primary brain tumors in adults, accounting for 12-15% of intracranial tumors and 50-60% of primary brain tumors [LHK03]. GBMs are highly malignant, infiltrate the brain extensively, and at times may become very large before turning symptomatic. The median survival time from the time of diagnosis without any treatment is 3 months, but with treatment survival of 1-2 years is common [VHN10].

Although the prognosis of GBM is uniformly poor, treating patients in an attempt to improve the quality of life is worthwhile.

GBM treatment consists of a combination of surgical resection, radiation therapy, and chemotherapy. Long-term disease-free environment is possible, but the tumor usually reappears, often within 3cm of the original site, and 10-20% may develop new lesions at distant sites. More extensive surgery and intense local treatment after recurrence has been associated with improvement.

Bevacizumab (Avastin) recently received FDA approval as a single agent for the treatment of patients with recurrent GBM following prior upfront, temozolomide (TMZ) - based chemoradiotherapy [Cha11]. GBM are highly vascularized cancers with elevated expression levels of vascular endothelial growth factor (VEGF), the dominant mediator of angiogenesis. Bevacizumab is a humanized monoclonal antibody that targets VEGF, and has been shown to improve patient outcomes in combination with chemotherapy (most commonly irinotecan) in recurrent GBM based on the positive results in two prospective phase II studies [Cha11].

1.2.2 Macdonald Criteria

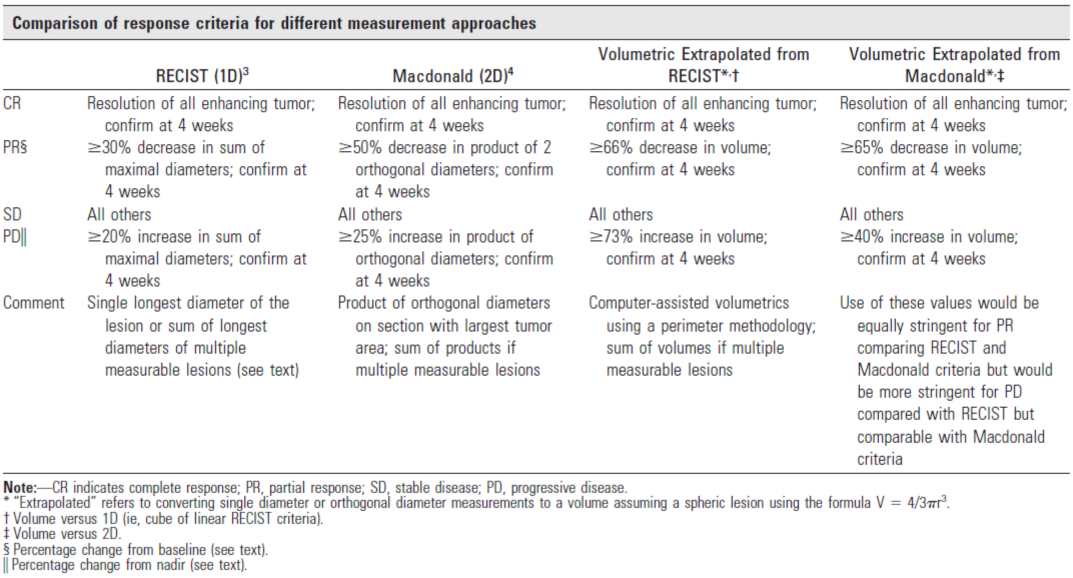
In the past three decades, great effort has been invested in clinical trials for malignant gliomas and brain metastases. Phase I and phase III studies, with their respective goals of defining maximal tolerated dose and overall survival, do not rely on neuroimaging as a primary end point. Imaging is crucial in phase II studies, because radiographic response, in combination with clinical status, is used to assess therapeutic effect. Phase II studies are usually conducted in patients with progressive tumors, and serial imaging examinations are performed after initiation of treatment and compared with a baseline pretreatment study. Radiographic response for each patient is then determined according to pre-determined criteria

[MCS90].

Radiographic response is often used as an end point in the phase II setting with the assumption that it is a valid surrogate measure for improved overall survival. After treatment, the response is described using four terms: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). Most trials for patients with malignant gliomas use the Macdonald criteria, as shown in Figure 1.1. The table compares 1D RECIST criteria and 2D Macdonald criteria.

The measures of response include best response (BR), time to progression (TTP), and progression free survival (PFS). BR is the largest reduction in tumor measurement during the study (compared with the initial measurements). TTP is the interval between the treatment start date and a subsequent imaging study that shows PD. PFS is the percentage of patients who have not experienced PD at a specified time point after beginning treatment (eg, 2-month PFS or 6-month PFS). A common end point for phase II clinical trials is 6-month PFS. Phase II studies compare the percentage of patients with PFS to that of a historical control group.

The definition of a measurable lesion is most important in the clinical trials that have an imaging end point. Contrast enhancement provides the best currently available measure of tumor size. The cystic and necrotic portions of the GBM tumor should be excluded, because they are unlikely to respond to interventions other than surgery. For multi-focal lesions, the approach is to measure each separate enhancing lesion and the sum the measurements. 1D and 2D diameter of the tumor are used for the evaluation of contrast-enhanced tumor size. Computer-aided volumetric methods are also under consideration, which can be more effective than diameters when the tumor contains a non-enhanced core or irregular shape.



Comparison of response criteria for different measurement approaches

	RECIST (1D)[3]	Macdonald (2D)[4]	Volumetric Extrapolated from RECIST*,†	Volumetric Extrapolated from Macdonald*,‡
CR	Resolution of all enhancing tumor; confirm at 4 weeks	Resolution of all enhancing tumor; confirm at 4 weeks	Resolution of all enhancing tumor; confirm at 4 weeks	Resolution of all enhancing tumor; confirm at 4 weeks
PR§	≥30% decrease in sum of maximal diameters; confirm at 4 weeks	≥50% decrease in product of 2 orthogonal diameters; confirm at 4 weeks	≥66% decrease in volume; confirm at 4 weeks	≥65% decrease in volume; confirm at 4 weeks
SD	All others	All others	All others	All others
PD‖	≥20% increase in sum of maximal diameters; confirm at 4 weeks	≥25% increase in product of orthogonal diameters; confirm at 4 weeks	≥73% increase in volume; confirm at 4 weeks	≥40% increase in volume; confirm at 4 weeks
Comment	Single longest diameter of the lesion or sum of longest diameters of multiple measurable lesions (see text)	Product of orthogonal diameters on section with largest tumor area; sum of products if multiple measurable lesions	Computer-assisted volumetrics using a perimeter methodology; sum of volumes if multiple measurable lesions	Use of these values would be equally stringent for PR comparing RECIST and Macdonald criteria but would be more stringent for PD compared with RECIST but comparable with Macdonald criteria

Note:—CR indicates complete response; PR, partial response; SD, stable disease; PD, progressive disease.
* "Extrapolated" refers to converting single diameter or orthogonal diameter measurements to a volume assuming a spheric lesion using the formula $V = 4/3\pi r^3$.
† Volume versus 1D (ie, cube of linear RECIST criteria).
‡ Volume versus 2D.
§ Percentage change from baseline (see text).
‖ Percentage change from nadir (see text).

Figure 1.1: Summary of Macdonald criteria - adapted from Macdonald et al. [MCS90]

1.2.3 RANO Criteria

Recently, limitations of the Macdonald criteria have been reported and new RANO criteria have been proposed [WMR10].

First of all, pseudoprogression could happen within three months due to radiotherapy effects, which limits the validity of PFS as an endpoint. To address this issue, the proposed new response criteria suggest that within the first 12 weeks of completion of radiotherapy, when pseudoprogression is most prevalent, progression can only be determined if the majority of the new enhancement is outside of the radiation field (for example, beyond the high-dose region or 80% isodose line) or if there is pathologic confirmation of progressive disease.

Secondly, increased enhancement often develops in the wall of the surgical cavity 48 to 72 hours after surgery, as a result of surgery and other therapies, not tumor recurrence. The RANO criteria propose that a baseline MRI scan should ideally be obtained within 24 to 48 hours after surgery and no later than 72

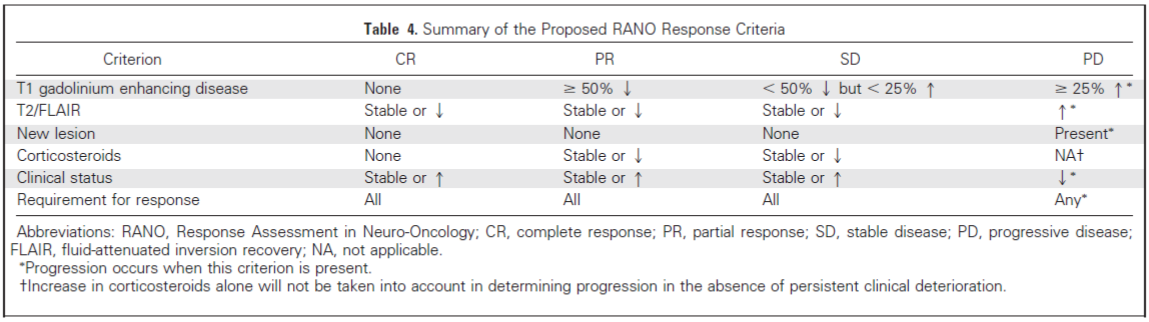


Table 4. Summary of the Proposed RANO Response Criteria

Criterion	CR	PR	SD	PD
T1 gadolinium enhancing disease	None	≥ 50% ↓	< 50% ↓ but < 25% ↑	≥ 25% ↑*
T2/FLAIR	Stable or ↓	Stable or ↓	Stable or ↓	↑*
New lesion	None	None	None	Present*
Corticosteroids	None	Stable or ↓	Stable or ↓	NA†
Clinical status	Stable or ↑	Stable or ↑	Stable or ↑	↓*
Requirement for response	All	All	All	Any*

Abbreviations: RANO, Response Assessment in Neuro-Oncology; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease; FLAIR, fluid-attenuated inversion recovery; NA, not applicable.
*Progression occurs when this criterion is present.
†Increase in corticosteroids alone will not be taken into account in determining progression in the absence of persistent clinical deterioration.

Figure 1.2: Summary of RANO criteria - adapted from Wen et al. [WMR10]

hours after surgery. The inclusion of diffusion weighted imaging in the immediate postoperative MRI scan can be helpful in determining whether new enhancement developing in the subsequent weeks or months is caused by tumor recurrence.

Thirdly, pseudoresponse could happen with antiangiogenic agents, especially those targeting VEGF. They can produce a marked decrease in contrast enhancement, as early as 1 to 2 days after initiation of therapy, which is not always necessarily indicative of a true antiglioma effect. As with the Macdonald criteria, the RANO criteria suggest that radiologic responses should persist for at least 4 weeks before they are considered true responses.

Lastly, the Macdonald criteria fail to measure non-enhancing tumor. RANO response assessment considers enlarging areas of nonenhancing tumor as evidence of tumor progression.

The RANO criteria are summarized in Figure 1.2.

1.2.4 Bi-dimensional or Volumetric Measurement?

Bi-dimensional measurements are defined as the product of the longest diameter and its longest perpendicular diameter, and is the methodology on which the WHO response criteria are based. There are a few limitations with bi-dimensional measurements.

Bi-dimensional tumor measurements are adequate surrogates for tumor volume only when tumors are spherical. However, such assumptions are often not true for GBM tumors, which tend to have irregular shapes, with necrosis or surgical cavities at their core. Obviously, the bi-dimensional measurement techniques cannot capture tumor changes along the z-axis, as the measurements are performed on a transverse image plane according to the criteria [ZSS09].

Studies [PID09, VUB03] have shown that there is substantial inter-reader variability in bi-dimensional measurements for brain tumors. Radiologists must subjectively select a single axial slice, which may differ from reader to reader. The bi-dimensional measurement on a single-axial image also ignores the tendency of malignant gliomas to be highly irregular in shape, to progress in a pattern of eccentric nodular growth, and to have cystic and central necrotic areas that are unlikely to be affected by non-surgical treatment [HUG08].

1.3 Survey of Automated Techniques in GBM Tumor Segmentation

In clinical studies, manual contouring has been used to segment tumors on MR images. For example, in a recent clinical study correlating Methylated-DNA-protein-cysteine methyltransferase (MGMT) promoter methylation and imaging features of GBM tumors, Drabycz et al. [DRR10] used manual contouring for GBM brain tumor segmentation. An accurate and robust automated segmentation system would facilitate quantitative analysis in clinical studies. In Figure 1.3, we show a 2D slice of a T1 weighted post-contrast MR image presenting an enhancing GBM brain tumor. On the right, we outline the active tumor region.

Automatic GBM brain tumor segmentation is a challenging task, since brain tumors are heterogenous, and highly variable in size, location, shape and appearance. They also often deform adjacent structures in the brain. Some artifacts of

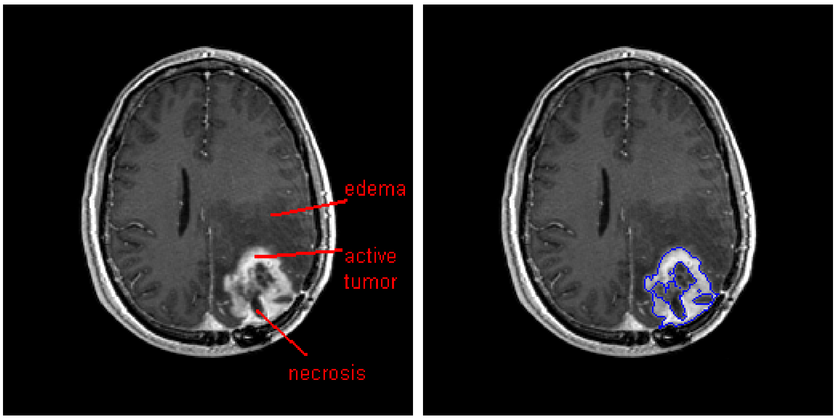


Figure 1.3: Example of the a GBM brain tumor on a T1w post contrast MR image slice and the corresponding tumor contour

MR imaging also increase the difficulty of tumor segmentation. The imperfection of the RF pulses and the location of RF coils may introduce non-uniformity in MR images. Our research focuses on recurrent GBM brain tumors that develop after surgery, many of which contain a cavity, and the enhancing portions can vary in shape, for example, ring-shape, blob-like shape, or multiple pieces attached to the cavity or dispersed into the brain tissues. Furthermore, when the patients were scanned at multiple centers, with different scanners and contrast agent injection protocols, the image intensity contrast can vary greatly among different scanners. All these factors makes GBM brain tumor segmentations a very challenging problem in a clinical setting, and there is a lack of previous studies evaluating GBM brain tumor segmentation methods in a large clinical dataset.

Computer-based brain tumor segmentation has remained largely experimental. Many efforts have exploited MRI's multi-dimensional data capability through multi-spectral analysis [PVP95, CHG98, HBG02, CSD08, DCY08, LUO05]. There are generally several categories of techniques: knowledge-based, clustering, voxel-based classification, level set methods, and graph-based techniques.

The knowledge-based segmentation systems typically use a brain atlas to provide prior information. Fletcher-Heath et al. [FHG01] applied a knowledge-based system to segment non-enhancing tumor. Prastawa [PBH04] applied outlier detection to find abnormal regions, k-means clustering (k=2) to separate tumor and edema, and then a region competition method, using level-sets to add smoothness constraints. The limitation is that they use T1-weighted pre-contrast and T2-weighted images, without contrast injection. But in clinical trials, tumor definition is based on T1-weighted post contrast images. They reported that the intra-reader consistency could be as low as 59.4% (overlap ratio between two sets of manual segmentations).

Among the clustering techniques, fuzzy clustering methods are the approach most widely employed across all tumor types. Fuzzy C-means (FCM) clustering is used frequently, since it does not require training data. Phillip [PVP95] was the first to apply the FCM clustering to GBM brain tumor segmentation, and correlated the segmentation with tumor histology. The limitation is that he did not give a quantitative validation of the method. Beevi et al. [BS10] applied efficient denoising algorithm before FCM and incorporated spatial probability to deal with the sensitivity to the noise. The limitation is that the method was validated on one clinical brain MR scan with unknown tumor type. Khotanlou et al. [KCA09] performed symmetry analysis and fuzzy clustering as an initialization segmentation, and combined deformable model and spatial relations to refine it. It is not clear whether the method was evaluated on GBM tumors, and it would be interesting to evaluate the method on images from GBM clinical trials. Aside from FCM, Ahmed et al. [AM08] performed K-means clustering combined with anisotropic diffusion denoising. The method was evaluated on only one MR scan, and further validation on more GBM tumors is needed. Liu et al. [LUO05] developed a semi-automated system using fuzzy-connectedness and evaluated the overall volume accuracies for 20 patients. The weakness is that additional efforts are need to remove attached brain structures. Clark [CHG98] developed a knowledge-based system including five stages, and knowledge is primarily T1-weighted, T2-weighted and PD-weighted image intensities. Each stage applies heuristic segmentation parameters. The reported performance has correspondence ratio ranging from 0.43-0.85 in 16 scans with 7 patients. They use 17 slices from 3 patients to set up the heuristic parameters. It is not clear whether it is possible to set universal parameter values in the setting of a large clinical trial, considering the variability of GBM tumors.

Voxel-based supervised classification methods have been investigated by a number of researchers [BHC93]. Vinitski et al. [VGK99] developed a system using a k-nearest neighbor classifier (kNN) to segment multiple sclerosis (MS) lesions and brain tumors from a limited number of patients. Validation with more tumor cases is needed to apply the method in clinical trials. Jolesz et al. [WKJ98] developed an adaptive template-moderated (ATM) classification algorithm (ATS) which incorporated a brain atlas to include spatial anatomical information into the kNN classification system and segment the MR image into five different tissue classes: background, skin, brain, ventricles, and tumor. Kaus et al. [KWN01] applied the algorithm to low-grade glioma and meningioma, however, it is not clear how the ATM algorithm will perform for GBM tumors. Prastawa et al. [PBM03] applied a system derived by Van Leemput et al. [VMV99] to GBM tumor segmentation. The system used the difference between T1w pre- and post-contrast to develop a tumor prior and an edema prior, and then form Gaussian Mixture Model framework solved by Expectation-Maximization (EM) technique, with an atlas prior as initialization to the EM. The performance is reported with an overlap ratio of 0.49-0.92 from 5 patients. The system was extended for GBM tumor segmentation by adding the tumor and edema classes [MBV02]. One limitation is that they did not provide a prior in the model for necrosis, cysts, or cavities. Another limitation is that the simplified geometric model for tumor shape cannot cope with tumors that have complex appearance and ambiguous boundaries.

Zhang [ZME04] used a baseline scan as training and follow-up scan as testing images. The method was tested on five scans from one tumor case. The application is limited since the GBM tumor on the baseline scan still needs to be manually contoured. Schmidt [SLG05] developed alignment based features: spatial prior, symmetry, intensity, multi-scale textural feature. The dataset has 10 patients with one cavitated tumor from two sites. They reported an average overlap ratio of 0.732. However, performance evaluation for active tumor volume is not clear. Lee [LSM05] applied a Discriminative Random Fields (MRFs) model with support vector machines (SVM) to 7 patients and reported performance of 0.53-0.89 overlap ratio for 12 scans from 7 patients. The weakness is that they used patient-specific training, which means training and testing voxels are from the same patient. This is not feasible in clinical practice since the manual contouring is still needed for each patient. Ayachi [AA09] applied support vector machine (SVM), using 9 slices from each tumor as training, and the rest of the slices on the same patient as testing, and reported 0.82 true positive rate for 4 cases. However, with patient-specific training, manual contouring is still needed. Zhang et al. [ZRL09] applied multi-kernel SVM, and again the limitation is patient-specific training.

Level set and graph-based methods were also explored for the brain tumor segmentation. Ho [HBG02] ran a level set on probabilities derived from T1w pre- and post-contrast difference. They reported 80-90% overlap ratio on 3 cases of blob-like shape. However, it is not clear how the method performs for irregular tumor shapes. Popuri et al. [PCJ09] extracted a clustered feature set, integrated them into a level set framework and used a Dirichlet prior to exclude the surrounding tissues. They showed success differentiating tumor from normal tissue by incorporating shape information, however, it is not clear how it performs for GBM tumors with irregular shapes. Taheri et al. [TOC10] used a threshold-based speed function for level-set function evolution. Corso et al. [CSD08] developed a segmentation by weighted aggregation (SWA) approach based on graph shift algorithm for GBM brain tumor segmentation. Dube et al. [DCY08] incorporated texture features into the SWA framework and applied it to GBM brain tumor segmentation on one-channel T1-weight post contrast MRI. The study achieved 70% accuracy for the majority of the cases, however, the failure cases need to be addressed before it is ready for the clinic. Recently, other features besides the intensity were studied, including grayscale concurrence matrix (GLCM) features [CBS09], discrete cosine transform (DCT) features [AMN10], and Gabor wavelets filter [Las10].

In summary, most of the literature uses multi-channel MR to segment GBM tumors, while segmentation on a single-channel MR has only been reported infrequently [DCY08]. Although multi-channel MR series are useful in differentiating brain tissues and disease, they are usually acquired at low resolutions, with slice gaps, and images from different sequences are often not aligned. Images can be realigned to a reference series but the re-sliced image series can suffer from lower resolutions along the slice axis as well as slice gaps. Segmentation on a single channel T1 post contrast isotropic data is potentially important in determining tumor volume for therapeutic response assessment in clinical trials.

Also, most of the above literatures used small datasets of less than 10 cases to evaluate their methods in segmenting GBM tumors. It is not clear whether they could handle the more difficult and irregular cases inevitably arising in larger datasets. There is a lack of previous studies evaluating GBM brain tumor segmentation methods in a large-scale clinical dataset.

Tumor recurrence could happen around a surgical cavity or at a distant site, and show diffuse-pattern with anti-VEGF drugs. These factors increase the difficulty of recurrent GBM tumor segmentation compared to newly-diagnosed GBM tumors.

Given that the accuracy of fully-automated method for GBM segmentation is currently not satisfactory, and the challenges of segmenting GBM tumors on a single-channel MR images, a meaningful contribution in clinical practice would be to develop a semi-automated system with minimal user interaction and high reproducibility. Our contribution is to build an interactive 3D segmentation tool, incorporating machine learned results to reduce the user interaction and improve reproducibility. To our knowledge, we were the first to improve semi-automated segmentation using machine learning with inter-patient training for GBM tumors. Our proposed framework has the potential to be generalized to other applications with appropriate training data. In this study, we applied the framework to a dataset of recurrent GBM tumors from a phase II trial, including 46 cases with heterogeneous tumor appearance. Both recurrent and non-recurrent tumors contain active tumor, necrosis and have irregular shapes, and thus the proposed framework can be applied to both tumor types.

1.4 Overview of Malignant Gliomas Characterization Using Imaging Features

In clinical trials, tumor size change is defined based on contrast enhancement on serial images. The limitations of the Macdonald Criteria become significant with the use of novel treatments. Both pseudoprogression, an increase in the nontumoral enhancing area, and pseudoresponse, a decrease in the enhancing area, show that enhancement by itself is not a measure of tumor activity but rather reflects a disturbed BBB [CRD11]. The most recent RANO criteria, therefore, suggest that the nonenhancing component of the tumor also be taken into account when making assessments of progression or response [WMR10].

Although the Macdonald Criteria have been widely used, false interpretations of tumor size increases on post-gadolinium-enhanced T1-weighted imaging may occur. New MR imaging and/or PET tools are needed to characterize tumors before initiation of therapy, depict the changes that result from treatment, and be validated as biomarkers of treatment effectiveness [CRD11]. The three available types of physiology-based MR imaging methods are: DWI, MR spectroscopy, and perfusion-weighted imaging. They were applied to either predict treatment response or accurately measure response of both enhancing and nonenhancing tumor shortly after treatment initiation. This could allow for earlier treatment decisions, saving patients from the adverse effects of ineffective therapies while allowing them to try alternative therapies sooner [Cha06].

To date, no single imaging technique has been validated to recognize and adequately establish a diagnosis. In spite of ongoing active research, the clinical utility of these physiologic imaging techniques remains unproven and the methods unstandardized. Much work lies ahead to validate and prove efficacy of these methods in improving diagnostic accuracy, affecting patient care, monitoring dynamic changes within brain tumor and normal brain during therapy [PYE11].

This section does not offer an exhaustive literature review, instead, a sample of the main imaging features that have been studied in the field of tumor characterization to monitor treatment response in the past few decades.

1.4.1 Anatomy-based Imaging Features

The characterization of brain tumors on anatomical images has been an active research area mainly in prognosis, tumor grading and disease detection. Morphological features such as the degree of necrosis and edema, the intensity of enhancement of the tumor, and the presence of large tumor cysts, were mostly studied as potentially significant prognostic imaging variables for malignant gliomas [HSS96, MSL04, PSP05]. Texture features enables quantification of gray-level patterns, pixel interrelationships, and spectral properties of images [KT10]. Researchers have mainly applied texture features for differentiation of disease and normal tissue, and for the classification of brain tumor type and grade.

However, for monitoring treatment response, few publications were found in the literature. It will be very interesting to investigate the significance of these features in treatment response assessment in the future.

1.4.2 Physiology-based Imaging Features

Physiology-based MR imaging methods have played a pivotal role in the transition of clinical MR imaging from a purely morphology-based discipline to one that combines structure with function. This section will overview the imaging markers that have been under investigation using four physiology-based MR modalities [Wal10].

1.4.2.1 MR Spectroscopy

Magnetic resonance spectroscopy (MRS) allows major metabolites to be measured in defined regions of tumors and the surrounding brain, notably choline-containing compounds (Cho, reflecting products of cell membrane turnover), N-acetyl aspartate (NAA, found in healthy neurons and axons), lactate (a product of anaerobic metabolism), myo-inositol (an astrocytic marker), and mobile lipid moieties (associated with necrosis).

Serial MRS has shown reduced choline levels in response to brachytherapy and gamma-knife radiotherapy and subsequent increases in Cho that can precede other markers of relapse [Wal10].

1.4.2.2 Perfusion Weighted MR

The use of perfusion imaging as a biomarker for response to antiangiogenic drugs has generated significant interest. There are no validated biomarkers for antiangiogenesis that are currently available for clinical use.

Perfusion weighted (PW) MR imaging techniques measure the degree of tumor angiogenesis and capillary permeability, both of which are important biological markers of malignancy, grading, and prognosis. The strength of PW-MR is the ability to depict changes in the internal architecture of the tumor in the setting of no overall change in tumor size. There are two most widely used techniques to quantify brain tumor vasculature - dynamic contrast ehanced (DCE) MR and dynamic susceptibility-contrast (DSC) MR.

From DCE-MR imaging, there are 2 primary end points according to a consensus recommendations and guidelines by a multi-disciplinary team [LBE10]: the volume transfer constant K_{trans} and the initial area under the gadolinium concentration-time curve (IAUGC) which are obtained on a voxel-by-voxel basis for assessment of antiangiogenic and antivascular therapeutics. From DSC-MR imaging, relative cerebral blood flow (rCBV) is the most widely evaluated for the assessment of therapeutic responses [PMB06].

Perfusion studies have shown longer survival in subjects with early decrease in rCBV one week into radiotherapy, and several studies have documented short-term increases and medium to long-term decreases in K_{trans} in response to radiation therapy in normal brain, gliomas, and meningiomas [Wal10].

As for Bevacizumab, Sawlani et al. [SRH10] found that percent change in hyperperfusion volume (HPV) (the fraction of tumor with an rCBV above 1.00) from baseline to first follow-up had a statistically significant hazard ratio of 1.07 when correlated with time to progression using 16 patients.

1.4.2.3 Diffusion Weighted MR

One of the most exciting potential applications of DW MR imaging has been in measurement of the response of solid tumors to therapy prior to measurable size change.

Increased tumoral ADC on DWI, measured at the 3rd week from the start of chemotherapy or radiotherapy, distinguished gliomas that subsequently showed partial response from those with stable or progressive disease [Wal10]. One group [MCM06a] developed a functional diffusion map (fDM) technique to take advantage of the relationship between ADC and cell density by examining voxel-wise changes in ADC measured in the same patient over time, and showed early changes in tumor diffusion values were highly correlated with patient survival. Ellingson et al. [EMR11] showed that Bevacizumab the rate of change of fDM-classified hyper-cellular regions (low ADC) within T2w abnormality regions is an early predictor of tumor progression, time to progression, and overall survival. These studies reported significant findings using DW images, however, GBM tumor biology after treatment is more complicated than being interpreted by only one factor of cell density. Edema and necrosis are significant components of recurrent GBM tumors, and these two factors show opposite effects in changing the tumor micro-environment. Thus, more sophisticated model is needed to represent the different factors with competing effects.

Pope et al. [PKH09b] applied a two Gaussian mixture model for ADC histogram analysis, and found that pretreatment ADC histogram analysis can stratify progression-free survival in Bevaizumab-treated patients with recurrent GBM. In this dissertation, the same two-component model will be applied to assess the therapeutic response.

1.4.2.4 Radiotracer

Radiotracers used in conjuction with positron emission tomography (PET) scans can measure tissue metabolism, and provide a potential biomarker in assessing gliomas. Chen et al. [CDS07] examined 19 patients treated with Bevacizumab and irinotecan. Response was defined as reduction in $^{18}F - FLT$ uptake by more than 25% after six weeks. This was significant predictor of overall survival.

1.5 Key Contributions and Outline

In this dissertation, we developed a computer-aided therapeutic response system (CADrx) for GBM brain tumors using T1w post-contrast MR and diffusion-weighted (DW) MR images in clinical trials. There are two main parts: 1) semi-automated segmentation of GBM brain tumors on T1w post-contrast MR; 2) early prediction of volumetric treatment response using ADC values derived from DW-MRI. The first part is the main focus of the dissertation.

1.5.1 Semi-automated Segmentation of GBM Tumors

Chapter 2 describes a sampling-based ensemble method to improve the reproducibility of a semi-automated method for GBM tumor segmentation. Lack of reproducibility and consistency is usually associated with semi-automated segmentation methods. In this study, we developed a new ensemble approach to improve reproducibility and applied it to GBM brain tumor segmentation on T1-weighted contrast enhanced MR volumes. The approach includes two novel steps: 1) Given a single user input, systematically generating a set of user input variations; 2) From all the user input variations, generating multiple segmentation results and applying ensemble method to obtain a final segmentation result. The reproducibility of the proposed framework was evaluated by a controlled experiment on 16 tumor cases from a multi-center drug trial. **The ensemble framework exhibited significantly better reproducibility than the standard semi-automated Otsu thresholding method.**

Chapter 3 further studies the potential of the ensemble framework in improving segmentation accuracy by aggregating different segmentation methods. There are two contributions: 1) We compared three segmentation methods on a relatively large dataset of GBM tumors (46 cases in our study vs 20 cases in the literature). 2) We showed that ensemble is not necessarily more accurate than individual methods. The final performance depends on the performance of the individual methods. In order for ensemble framework to outperform, the individual methods have to be complimentary to each other. Based on the study of Chapter 3, we invented a novel system to ensemble semi-automated segmentation method and fully-automated voxel classification using machine learning methods, as described in Chapter 4. **To our knowledge, we were the first to validate the methods on a large clinical dataset. The heterogeneous nature of GBM tumors suggests that an ensemble method may be appropriate.**

In Chapter 4, we developed a novel system to to build an interactive 3D segmentation tool, incorporating machine learned results to reduce user interaction and improve the reproducibility. There are two contributions: 1) Incorporating machine learning into the interactive segmentation framework. 2) The classifier is trained with focused sampling strategy, instead of conventional random sampling. Our proposed framework has the potential to be generalized to other applications. **To our knowledge, we are the first to improve semi-automated segmentation using machine learning with inter-patient training for GBM tumors.**

1.5.2 Early Prediction of Treatment Response Using ADC Values

Chapter 5 develops a quantitative quality control (QC) method for ADC values obtained from multiple centers. With our limited dataset, we developed a tool to evaluate the variability of ADC measurement across different sites and different scanners. The tool draws a fixed-size region of interest (ROI) on the normal appearing brain white matter, and calculates the mean and standard deviation (STD) of ADC values in the ROI. The tool provides quantitative quality control (QC). Future work is to validate the tool with phantom and healthy volunteer data.

Chapter 6 develops a computer-aided therapeutic response assessment (CADrx) system for early prediction of drug treatment response for GBM brain tumors with DW MR images. In conventional Macdonald assessment, tumor response is assessed nine weeks or more post-treatment. This study used DWI at the 5th week to predict the volumetric change at the 9th week. The contribution of this chapter is to run multi-parametric analysis and apply machine learning methods to classify responders vs non-responders. The feature set includes descriptive statistical features, Earth Mover's Distance (EMD) to measure the difference between two histograms, and two-modal Gaussian Mixture parameters. **To our knowledge, we were the first to publish the potential of multiple ADC parameters using machine learning methods to classify response in GBM.**

CHAPTER 2

Sampling-based Ensemble Segmentation against Inter-operator Variability

Abstract

Inconsistency and irreproducibility are commonly associated with semi-automated segmentation methods. In this study, we developed an ensemble framework to improve the reproducibility and applied to GBM brain tumor segmentation on T1-weighted contrast enhanced MR volumes. The proposed approach combines sampling-based simulations and ensemble segmentation into a single framework; it generates a set of segmentations by perturbing user initialization and user-specified internal parameters, then integrates the set of segmentations into a single consensus result. Three combination algorithms are applied: majority voting, averaging and expectation-maximization (EM). The reproducibility of the proposed framework was evaluated by a controlled experiment on 16 tumor cases from a multi-center drug trial. The ensemble framework had significantly better reproducibility than the individual base Otsu thresholding method ($p < .001$).

2.1 Introduction

In clinical practice, reproducible and repeatable segmentation is an important prerequisite for longitudinal study of medical images. Semi-automated segmentation methods are often preferred in common radiographic protocols because it allows expert clinicians to control the segmentation quality which plays a critical role in the final diagnostic decision. However, such semi-automated methods are also inconsistent by design when administered by different readers and/or used with different internal parameter values. This is a trade-off between usability and repeatability, posing a serious technical challenge. Our study focuses on segmenting GBM brain tumors in T1-weigted contrast enhanced MR volumes by using the Otsu thresholding method. GBM tumor segmentation offers an ideal test case for our study because of the contrast-enhancing heterogeneity of the tumors makes the current state-of-the-art methods highly irreproducible. Otsu method is used in this study and has its advantage in efficiency, simplicity and usability, however, it also suffers from poor reproducibility due to its semi-automated design and inhomogeneous tumor nature. Different user interaction will lead to different thresholds and thus inconsistent segmentation results. The algorithm parameter setup, here the number of thresholding levels, will also result inconsistent segmentation results.

The main purpose of our study is to improve the reproducibility and stability of the semi-automated Otsu thresholding segmentation applied to GBM brain tumors. The final goal is to generate a robust and stable ensemble result given one single manually-drawn user interaction. The proposed framework combines sampling-based simulations and ensemble segmentation into a single framework; it generates a set of segmentations by sampling user interaction space and algorithm internal parameter space, then the set of inconsistent segmentations are ensembled into a single consensus result. The algorithmic sampling of user interaction space is designed to perturb the manual user interaction to simulate the typical inter-operator variability, thus the fusion of all possible segmentations is expected to be stable, reproducible and repeatable.

2.2 Materials and Methods

The brain volumes were pre-processed by skull-stripping using FSL tools [SJW04]. As a pre-process, users define a bounding cube as the volume of interest (VOI), and all the segmentation is done within the VOI. The user interaction is to draw a 2D bounding box around the tumor on a 2D slice.

The base segmentation algorithm is Otsu thresholding method [Ots79] because of its efficiency, however, it suffers from poor reproducibility due to its semi-automated design and inhomogeneous tumor nature. The concept of Otsus thresholding method is to find the intensity threshold value that minimizes the weighted within-class variation: $\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_1(t)\sigma_2^2(t)$, with i as each intensity bin in the histogram $p(i)$, the class probability defined as $q_l(t) = \sum_{i=1}^{t} p(i)$, class variance as $\sigma_l^2(t) = \sum_{i=1}^{t}[i - \mu_l(t)]^2 \frac{p(i)}{q_l(t)}$, class mean as $\mu_l(t) = \sum_{i=1}^{t} \frac{ip(i)}{q_l(t)}$, and $l = 1, 2$ as two classes. Given an initial $\mu_l(0)$ and $q_l(0)$, the algorithm will do an exhaustive search by altering the thresholding value t to find the optimal thresholding value $t_{opt} = argmax\ \ \sigma_w^2(t)$. The idea can be generalized to multiple classes.

The segmentation process using the base method is described as follows. First, given a 2D bounding box around the tumor, intensity values within the box are collected to form a histogram $p(i)$, and Ostu thresholding method is applied to find the optimal threshold t_{opt}, with the assumption that the number of classes within the bounding box is $L = 2$ or $L = 3$. Then, the 3D VOI is thresholded with the highest thresholding value, followed by successive application of a connected component analysis, morphological opening and closing. The structuring element used for morphological operations is [0,1,0;1,1,1;0,1,0].

In this study, we proposed an ensemble segmentation framework, and the goal is to generate a robust and stable ensemble result given one manually-drawn 2D box. There are two steps in the proposed ensemble framework: 1) a set of inconsistent segmentations are generated by sampling user interactions and

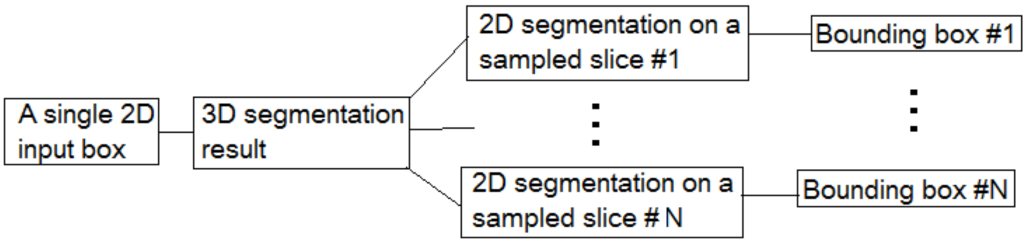

A single 2D input box
3D segmentation result
2D segmentation on a sampled slice #1
2D segmentation on a sampled slice # N
Bounding box #1
Bounding box #N

Figure 2.1: Pipeline of automated sampling of user inputs to simulate inter-operator variability

algorithm internal parameters; and 2) the set of inconsistent segmentations are fused into a single consensus result.

2.2.1 Sampling User Interactions

The purpose of this step is to automatically sample a set of 2D bounding boxes on different slices that simulate typical inter-operator variability, given one 2D bounding box drawn by the user manually on one single slice.

First, given the manually-drawn 2D input box, a 3D segmentation result is generated using the base method. Among all the slices occupied by the 3D segmentation, N slices are sampled uniformly. On each of the N slices, one 2D box of the tumor is generated by acquiring the bounding box of the tumor contour. In the end, the set of N 2D bounding boxes is generated as the user interaction variation set. This sampling pipeline is shown in Figure 2.1.

2.2.2 Sampling Internal Algorithm Parameters

The internal parameter to for Otsu method is the number of intensity thresholding levels within a user input box. The difficulty of this parameter setup comes from the GBM tumor appearance. GBM brain tumors may be composed of various tissue types, such as necrosis, tumor cells and edema, which present different

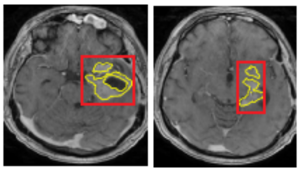


Figure 2.2: Two slices of one tumor: left: 3-object problem; right: 2-object problem

levels of enhancement, but not always present within the same tumor. Figure 2.2 shows two example cases of $L = 3$ (left) and $L = 2$ (right). Apparently it is not accurate to set a universal number of levels for all 2D user interaction boxes. Thus, we run the base method with both $L = 2$ and $L = 3$ in this study.

2.2.3 Ensemble Segmentation

The automatically sampled N user interaction boxes in Figure 2.1 are used to perform independent 3D Otsu segmentations, and two intensity levels ($L = 2$ and $L = 3$) are applied to each user interaction box. As a result, 2N inconsistent 3D segmentation results are generated. In this step, the 2N inconsistent segmentation results will be fused into a final ensemble segmentation result.

We compared three ensemble methods: 1) majority voting [KHD98], 2) averaging and 3) STAPLE-EM algorithm [WZW04], to fuse the 2N inconsistent segmentation results. The averaging rule is defined as follows: $PM(i, j, k) = \frac{1}{N}\sum_{n=1}^{N} Seg_n(i, j, k)$, with PM as the probability map (PM) and Seg_n as the binary segmentation for n-th individual segmentation result for each voxels (i, j, k). The majority voting rule is to threshold the PM with 0.5 into a binary segmentation. For the details of STAPLE-EM, readers are referred to [WZW04]. Both averaging and EM method generate a probability map, which is thresholded into binary segmentations by 0.3. This value of 0.3 was determined empirically.

In a summary, the ensemble framework takes a single manual 2D box as the

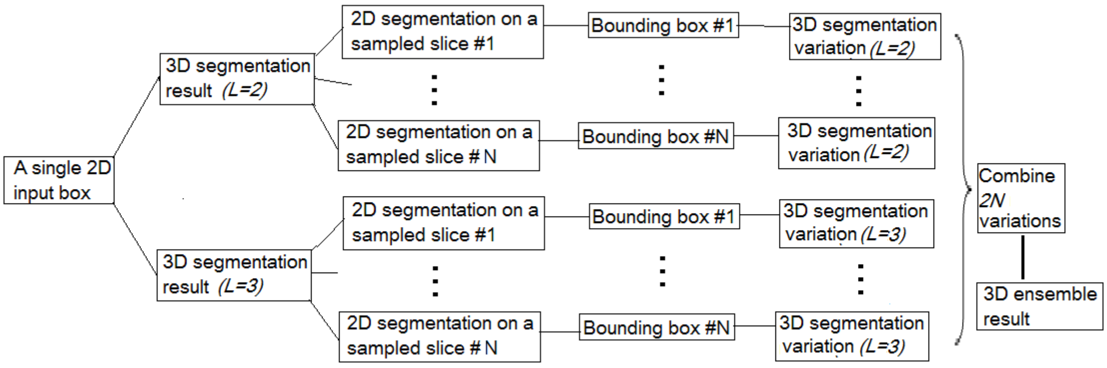

A single 2D input box
3D segmentation result (L=2)
3D segmentation result (L=3)
2D segmentation on a sampled slice #1
2D segmentation on a sampled slice # N
Bounding box #1
Bounding box #N
3D segmentation variation (L=2)
3D segmentation variation (L=3)
Combine 2N variations
3D ensemble result

Figure 2.3: The complete ensemble segmentation pipeline

input, and generates one ensemble 3D segmentation as the output. The complete framework is shown in Figure 2.3. In this study, we set N=8.

2.3 Experiments and Results

We randomly selected 16 GBM tumor cases from our research database with T1-weighted post-contrast volumetric MR images (voxel size 0.9*0.9*1mm).

The reproducibility of the proposed ensemble framework was evaluated by a controlled experiment. For each tumor, one lab technician manually drew eight 2D bounding boxes on 8 uniformly-sampled slices across the whole tumor volume. For each of the manually-drawn box, the base method and the ensemble framework were applied respectively for comparison. The overlap ratio of the 8 base Otsu results and that of 8 3D ensemble results were calculated and compared. Overlap ratio is defined as $OR = \frac{A \cap B}{A \cup B}$, with A and B as binary segmentation results. Figure 2.4 compares the individual results and ensemble result with averaging rule. The averaging ensemble method showed significantly better reproducibility than the base Otsu method with $L = 3$ ($p < .001$) and $L = 2$ ($p < .001$). Figure2.5 compared the three ensemble rules. Although there was no statistically significant difference with this dataset ($p > 0.05$), STAPLE-EM was slightly better

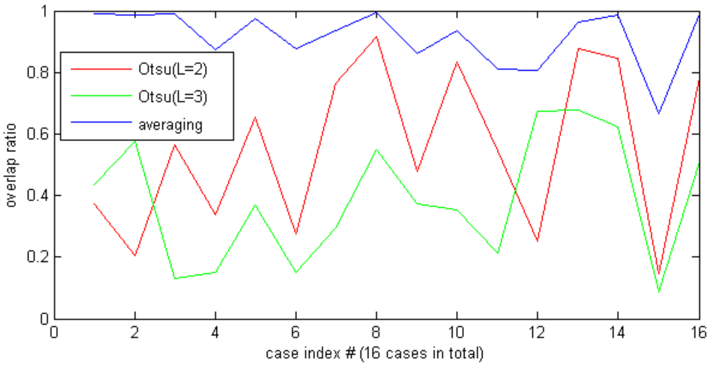


Figure 2.4: Reproducibility comparison between individual Otsu and averaging ensemble

than averaging and voting visually.

The accuracy of the proposed framework was evaluated by calculating F_1-measure [Rij79] between the ground truth and the semi-automated methods.

$$F_1{-}measure = 2*\frac{precision * recall}{precision + recall},\ with\ precision = \frac{tp}{tp+fp},\ recall = \frac{tp}{tp+fn}$$

, with tp as true positive, fp as false positive, and fn as false negative. For our dataset, with eight manually-drawn boxes for each of the 16 tumor cases, there were in total 128 boxes. For each user interaction box, the accuracy was shown in Figure 2.6. All three ensemble methods showed significant improvement over the Otsu method with $L = 3$ ($p < .001$), but no difference from Otsu method with $L = 2$ ($p > .05$). There was no statistically significant difference between the three ensemble methods in accuracy ($p > .05$) as shown in Figure 2.7.

Figure 2.8 and Figure 2.9 showed an example comparing the segmentation results using different manually-drawn user interaction boxes between the base Otsu method and the ensemble framework with the averaging rule. Different rows show different manually-drawn user interaction boxes; while columns show

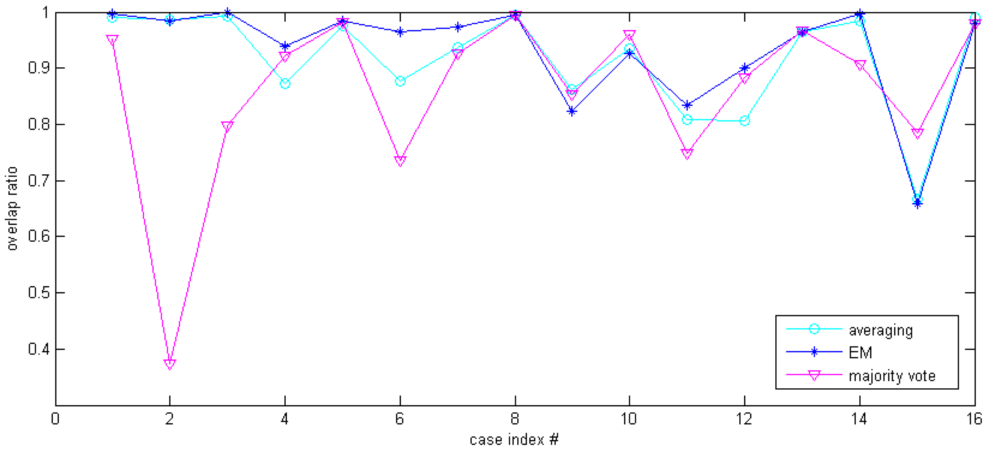


Figure 2.5: Reproducibility comparison between the three ensemble rules

different slices of the same tumor. Figure 2.8 demonstrated that the base Otsu segmentations are inconsistent using different manually-drawn boxes. Figure 2.9 illustrated that the ensemble segmentations are consistent. Even at the presence of the user interaction variability, the similarity of results within each column indicates the consistency.

2.4 Discussion and Conclusion

We developed an ensemble framework aiming to improve the reproducibility of the semi-automated segmentation, and applied the framework to GBM brain tumor segmentation on T1w post contrast images. First, we invented an automated sampling of user interactions that simulated typical inter-operator variability, as well as sampling internal algorithm parameters. Then, we generated the inconsistent segmentation set using the sampled user interaction boxes and ensemble them into a final segmentation. We evaluated the performance on a difficult task of the single-channel GBM brain tumor segmentation.

Ensemble results were worse than the individual results for the first two cases as

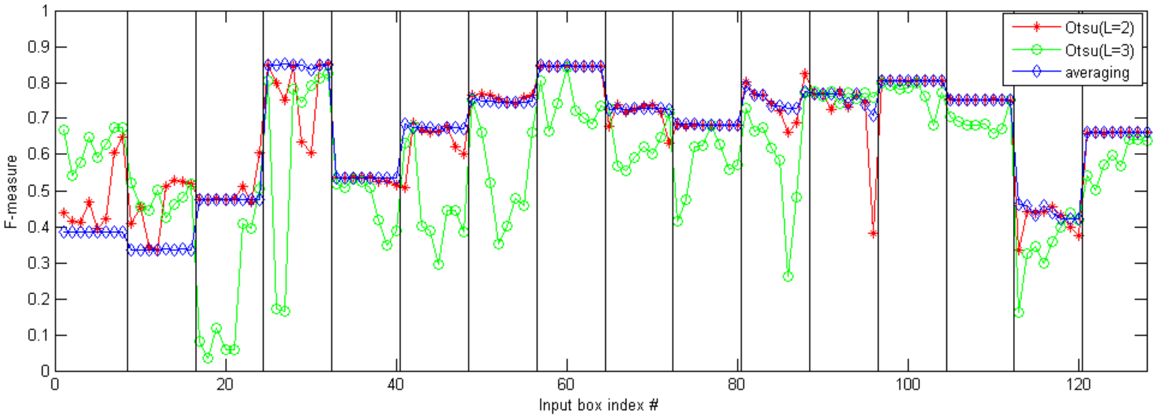


Figure 2.6: F-measure of the individual methods and the averaging ensemble

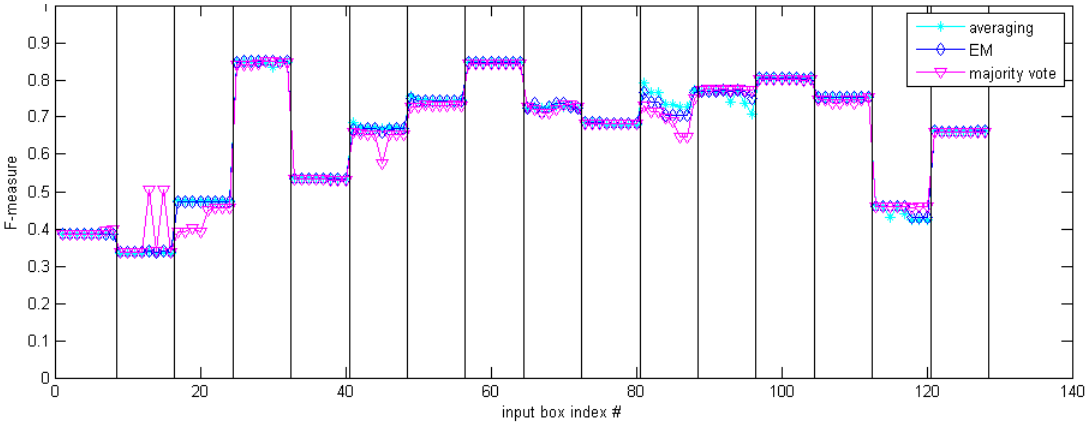


Figure 2.7: F-measure of the three ensemble methods

shown in Figure 2.6, because the majority of the individual results did not include the ”fuzzy” part of the tumor when the contrast enhancement is heterogeneous. For cases No. 4, 6, 12, and 15, it is clearly shown that ensemble is more consistent and accurate than the individual segmentation results.

The limitation of the study is that generating multiple segmentations is computationally expensive, thus the proposed system has limited feasibility to be applied in clinical practice. To overcome this problem, parallel computing can be applied to run multiple segmentation algorithms, and can be realized by graphics processing unit (GPU).

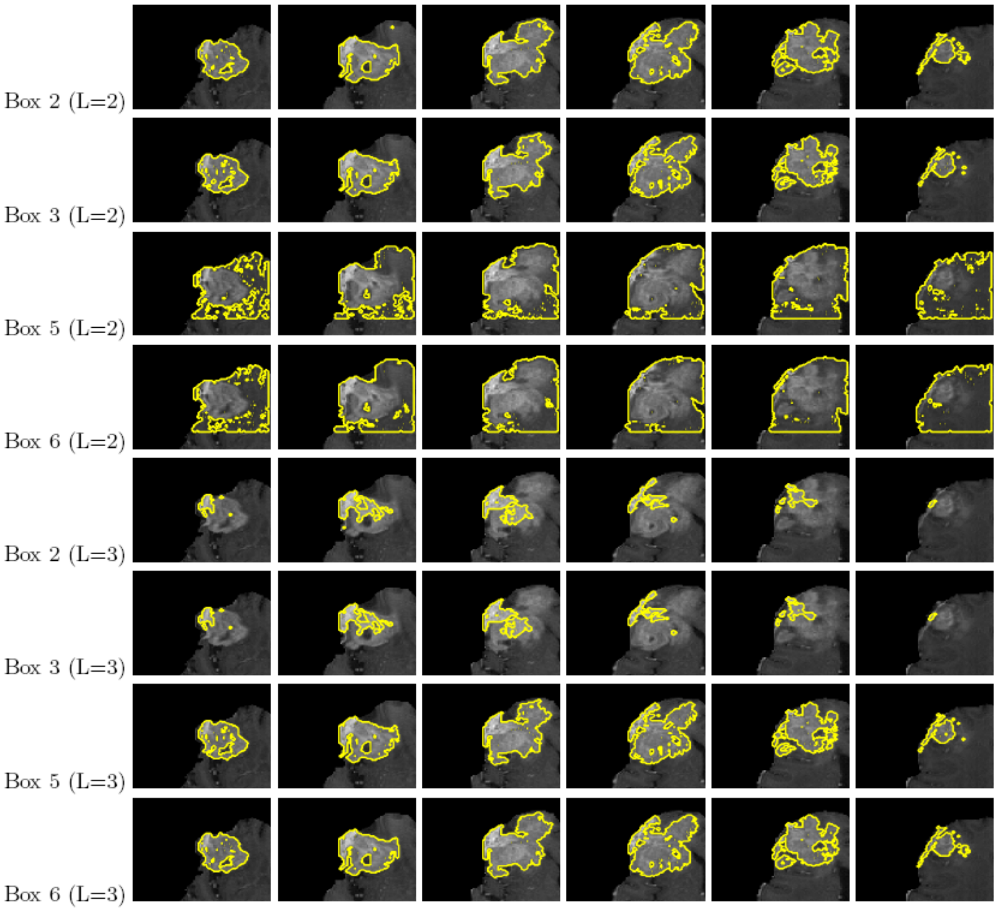


Figure 2.8: Segmentation results from the base Otsu method using different manually-drawn user interactions. Columns are 6 different slices of the same tumor; rows are result from different manually-drawn user interaction boxes

In conclusion, the proposed automated user interaction sampling and ensemble framework significantly improved the reproducibility compared to the base method on our dataset on GBM brain tumor segmentation with comparable overall accuracy. Reproducibility is crucial for semi-automated segmentation methods, the proposed framework shows potential in improving the consistency and reproducibility.

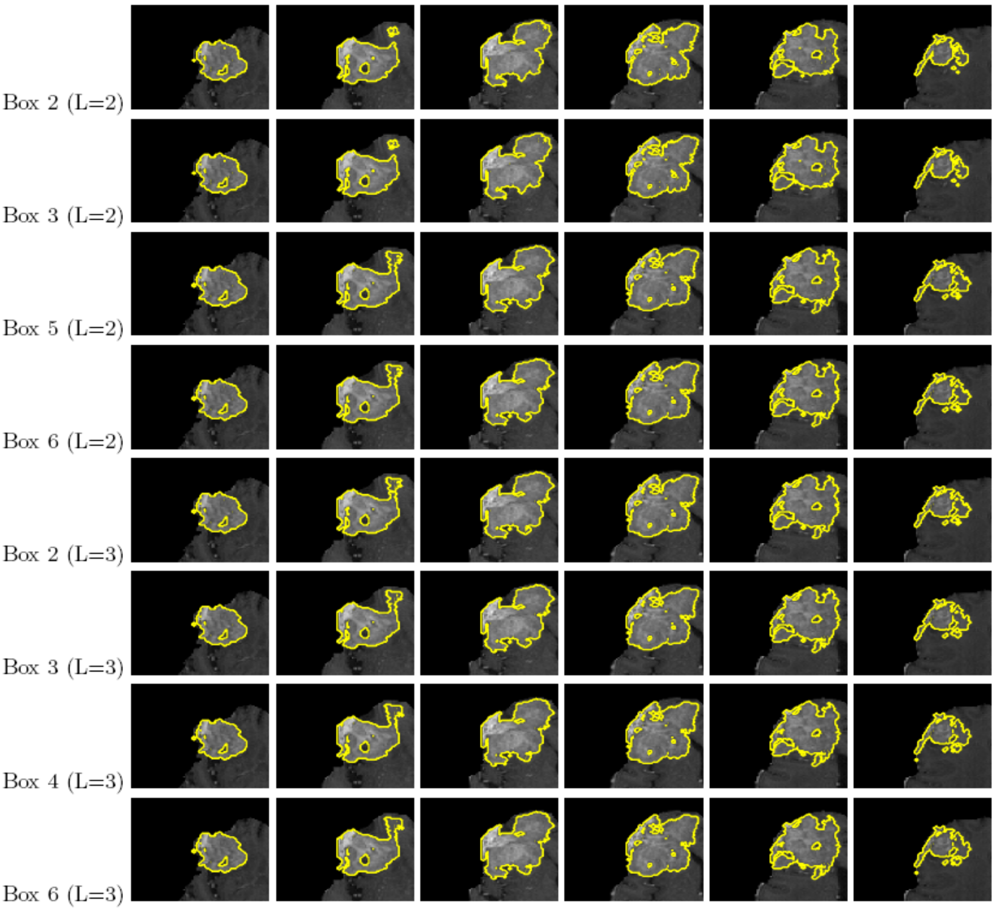


Figure 2.9: Segmentation results from the ensemble framework using different manually-drawn user interactions. Columns are 6 different slices of the same tumor; rows are result from different manually-drawn user interaction boxes

CHAPTER 3

Ensemble Segmentation for GBM Brain Tumors on MR Images Using Confidence-based Averaging

Abstract Typically there exists no single segmentation method that outperforms others for all cases in a given application domain. Ensemble segmentation methods run multiple algorithms and combine the results with the goal of achieving better robustness and accuracy. The goal and contribution of this study is to develop an ensemble segmentation framework for GBM tumors on single-channel T1w post-contrast MR images, and evaluate the performance on a relatively large dataset of 46 subjects including different types of tumor appearances in a pharmaceutical drug trial. Three base methods were evaluated in the framework: fuzzy connectedness, GrowCut, and voxel classification using support vector machine. A confidence map averaging (CMA) method is used as the ensemble rule. The results showed that the CMA ensemble is consistently close to the best performed base method for each case.

3.1 Introduction

In this study, we investigate an ensemble approach to GBM tumor segmentation that combines results from three general-purpose segmentation algorithms, aiming to achieve high accuracy in GBM tumor segmentation while maintaining the generalizability to other applications.

There has been active research on combining multiple segmentation results. In the field of supervised learning, Kittler [KHD98] summarized the different schemes for combining results from multiple classifiers. In the field of unsupervised clustering, Ghaemi [GSI09] performed a survey of methods in clustering ensembles. As far as applications in the medical imaging field, Wattuya et al. [Gra06, WRP08] developed an algorithm to combine multiple segmentation results using the random walker method. Rohlfing et al. [RRM04] studied atlas-based segmentation of biomedical images. They proposed to estimate the performances of the base classifiers and combine their respective outputs by weighting them according to their estimated performance. This method is a multiclass extension of an EM algorithm for ground truth estimation from a binary classification based on decisions of multiple experts [WZW04]. Aljabar et al. [HHA06, AHH09] applied a majority voting rule [KHD98] to combine segmentation results from an atlas-based segmentation and presented a thorough evaluation on brain MR images.

In this study, we propose an ensemble technique, applied to semi-automated GBM brain tumor segmentation on T1w post-contrast volumetric MR images, and evaluate the performance on a dataset with 46 tumor cases from a clinical trial research database. There are two steps involved. The first step is to generate input segmentation candidates from different algorithms. Three general-purpose segmentation methods were applied to generate input segmentations: fuzzy connectedness [LUO05], GrowCut [VK05], and voxel classification using support vector machines (SVM) [DHS00]. The second step is to combine them to generate a final result. The ensemble scheme was confidence-based averaging (CMA). The CMA method was adopted based on the assumption that the majority of the base methods are correct, and errors from each method are independent so that they will be averaged out in the ensemble result. To our knowledge, we are the first to investigate ensemble segmentation for GBM tumor segmentation on single-channel MR images (T1w post contrast), and to evaluate base methods and their ensemble on a dataset of 46 GBM tumors including different types of GBM tumor appearance patterns. Previous studies reported in the literature have used smaller datasets of 5-20 cases.

3.2 Materials and Methods

3.2.1 Input Segmentations

We explored three algorithms as base methods including two semi-automated methods and one learning-based: fuzzy connectedness, GrowCut, and voxel classification using SVM. The fuzzy connectedness method was selected because it was validated to work well for semi-automated GBM brain tumor segmentation by Liu et al. [LUO05]. The GrowCut method was chosen due to its simple user interaction, straightforward implementation, and promising performance in our pilot study [HRO11]. The learning-based segmentation method was chosen here so that the general-purpose method can be adopted to this specific application by learning from examples.

3.2.1.1 Fuzzy Connectedness

The Fuzzy Connectedness (FC) segmentation framework assigns fuzzy affinities to the target object during segmentation. The fuzzy connectedness captures global fuzzy “hanging togetherness”. In practice, the first step computes an “affinity” map, a local fuzzy relation, which quantifies the connectedness of any pixel pair in the original image; the second step calculates the “fuzzy connectedness”, the global fuzzy relation with one specific (designated) pixel belonging to the object of interest.

We implemented the algorithm following Liu’s work [LUO05] since it has been previously applied to the GBM brain tumor segmentation task. First, the affinity between any two voxels c and d, denoted by $\mu_k(c, d)$, is given by:

$$\mu_k = \begin{cases} 1 & \text{if } c{=}d \\ 0 & \text{if } c \text{ and } d \text{ are not 6-adjacent} \\ h_1(f(c), f(d)) * h_2(f(c), f(d)) & \text{otherwise} \end{cases}$$

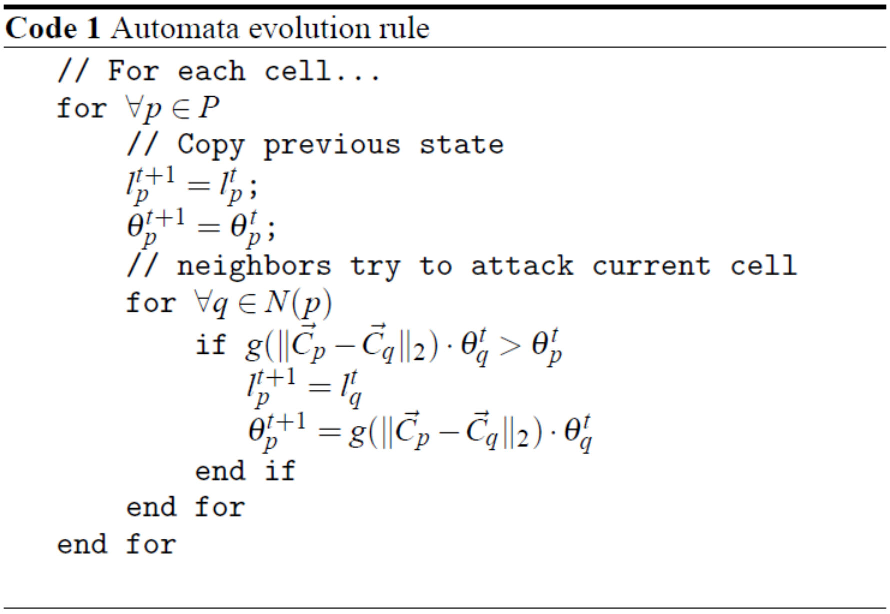
where $f(c)$ and $f(d)$ denote voxel intensity values at c and d, respectively.

The functional forms for h_1 and h_2 are chosen as follows,

$$h_1(f(c), f(d)) = exp\left((-1/2)\left[\left(\left|\frac{f(c)-f(d)}{f(c)+f(d)}\right| - m_1\right)/s_1\right]\right)$$

$$h_2(f(c), f(d)) = \begin{cases} 0 & \text{if } f(c) + f(d) < a1 \\ \frac{f(c)+f(d)-a1}{a2-a1} & \text{if } a1 < f(c) + f(d) < a2 \\ 1 & \text{if } f(c) + f(d) > a2 \end{cases}$$

m_1 is set to the mean of the relative intensity differences $|f(c) - f(d)/(f(c) + f(d)|$ computed for all 6-adjacent voxel pairs (c, d) within the region. s_1 is set to two times the standard deviation of this relative difference within the user input seeds. a_1 is set to [mean$-$two times the standard deviation of intensity sums $f(c)+f(d)$ of all 6-adjacent voxel pairs (c, d) within the region]. a_2 is set to [mean + (two times the standard deviation of intensity sums $f(c) + f(d)$ of all 6-adjacent voxel pairs (c, d) within the region]. Second, the strength of the fuzzy connectedness is calculated by dynamic programming. There are numerous paths between any two given voxels c and d. In each possible path, the “strength of connectedness” is simply the smallest pairwise neighboring fuzzy affinity along this path. Among all possible paths, the one with the largest strength is the fuzzy connectedness of the two voxels c and d. In the end, the pool O of voxels with non zero membership



Code 1 Automata evolution rule

```
    // For each cell...
    for $\forall p \in P$
        // Copy previous state
        $l_p^{t+1} = l_p^t$;
        $\theta_p^{t+1} = \theta_p^t$;
        // neighbors try to attack current cell
        for $\forall q \in N(p)$
            if $g(\|\vec{C}_p - \vec{C}_q\|_2) \cdot \theta_q^t > \theta_p^t$
                $l_p^{t+1} = l_q^t$
                $\theta_p^{t+1} = g(\|\vec{C}_p - \vec{C}_q\|_2) \cdot \theta_q^t$
            end if
        end for
    end for
```

Figure 3.1: Pseudo code of the cellular automata evolution rule - adapted from [VK05]

value in the fuzzy subset satisfies all of the following conditions: (1) all seed voxels are in O; (2) for any two voxels c and d in O, their strength of connectedness $S(c, d) > \theta$; (3) for any voxels c in O and d not in O, $S(c, d) < \theta$.

3.2.1.2 GrowCut

The GrowCut method [VK05] (GC) is based on cellular automata theory. Formally, a cellular automaton (CA) is a triple (S, N, δ), where S is the state set, N is the neighborhood, and $\delta : \ S^N -> \ S$ is the local transition function, where S^N indicates the states of the neighborhood cells at a given time, while S is the state of the central cell at the next time step. In the GrowCut method, the cells correspond to image voxels, and the cell state $S = (C, l, \theta)$ for each voxel consists of the image feature vector C, typically the voxel intensity, the label l indicating which category the voxel belongs to, and the strength θ in the continuous range $[0, 1]$ representing the confidence in the current labeling.

The GrowCut method uses CA theory to interactively label the image volume using user supplied seeds. The user starts the segmentation by specifying the seeds on both tumor and background voxels, the seeds' labels are set to the respective category labels, and their strength is set to 1. This sets the initial state of the cellular automaton. Strengths for unlabeled cells are set to 0. In each iteration t, each cell tries to "attack" its neighboring voxels by calculating the local intensity similarity; accordingly, the label map and the strength map are updated until convergence. The algorithm converges to a stable configuration, where no cells change state. The pseudo code for the GrowCut algorithm is shown in Figure 4.4, where $N(p)$ is 26-neighborhood system of a voxel p in 3D, and g is a monotonous decreasing function bounded to [0,1],

$$g(x) = 1 - \frac{x}{max||C||_2}$$

3.2.1.3 Voxel Classification Using SVM

The support vector machine (SVM) [DHS00] is a supervised learning algorithm. The SVM constructs a separating hyperplane in a N-dimensional space where class data can be viewed as sets of feature vectors, and the hyperplane maximizes the margin between the data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the data sets. Considering a two-class problem, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. Given a set of n labeled data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ where $y_i = \pm 1$ and x_i is the feature vector, and SVM searches for a optimal separating hyperplane $< w, x > + b = 0$, where where

$w \in R^n$, $x \in R^n$, and $b \in R$. The object function to minimize is as follows:

$$\frac{1}{2} * ||w||^2, \ subject \ to \ y_i(< w, x_i > \ + b) \geq 1$$

The optimization problem can be solved by Lagrange multipliers method. This method introduces an unknown scalar variable α_i for each constraint and forms the object function as follows:

$$f(x) = \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j$$

It often happens that the sets to be discriminated are not linearly separable in the original feature space. For this reason, SVM can map the data into a high dimensional nonlinear feature space, and construct an optimal linear hyperplane in this space. This mapping is performed by the kernel function $\varphi(x)$. The object function can be further formed as:

$$f(x) = \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i) \varphi(x_j)$$

One common kernel function $\varphi(x)$ is radial basis function:

$$\varphi(x, y) = exp(-\gamma ||x - y||^2)$$

There are two stages classifying image voxels with supervised learning: a training stage in which the system is developed using the ground truth voxels, and a testing stage where the system is applied to unknown voxels.

In the training stage, voxels from manually contoured tumors were used as positive (tumor) examples, and an equal amount of voxels sampled outside the tumor were used as negative (background) examples. For each training sample, a set of imaging-based features was calculated: intensity, gradient magnitude, first-order Gaussian derivatives (in three directions), second-order Gaussian derivatives (six in total), and the three eigenvalues of the Hessian matrix. These features are calculated on three different scales - 1, 2, and 4 pixels. In total, we have 42 features derived from images as the feature vector to train the SVM classifier.

To apply the voxel classification to the test scan, the set of 42 features is calculated for each voxel and fed into the trained classifier, and for each voxel a probability that it belongs to a tumor will be assigned.

3.2.2 Combining Input Segmentations

Confidence maps (CM) is defined for the labeling by each base method. For the SVM method, the output probability map was used as the CM. For the Grow-Cut method, a strength map was generated by the algorithm, and we transform the strength map into the confidence map by linearly re-scaling the foreground strength to [0.5,1] and the background strength to [0, 0.5]. For the fuzzy connectedness method, the membership value is linearly re-scaled to [0,1] as the CM.

The three base methods (N=3) are combined by confidence map averaging (CMA). The output of the ensemble is the average of the three confidence maps generated by the three base methods, weighting each of the three methods equally. In order to obtain the binary segmentation, the CMA result is later thresholded to obtain a binary segmentation.

$$CMA(i,j,k) = \frac{1}{N}\sum_{n=1}^{N} CM_n(i,j,k)$$

3.3 Experiments

We have 46 GBM tumors from 45 patients in this study from a 60-subject multicenter clinical trial database. The 15 patients were excluded due to either lack of manual tumor drawing, anisotropy of voxel size, or variation in image resolution.

The imaging protocol for T1w is 3D volumetric acquisition in the axial plane with flip angle-spoiled gradient echo sequence (FSPGR) or magnetization-prepared rapid gradient-echo (MP-RAGE) sequence with 1mm slice thickness, 0.9mm by 0.9mm pixel size, and 256*256 in-plane resolution.

The ground truth for the segmentation was manually contouring by a board-certified neuroradiologist with 10 years of experience, facilitated by a semi-automated segmentation tool [Ots79] from an in-house software system QIWS (Quantitative Imaging Work Station).

The brain volume was preprocessed to remove non-brain matter and obtain consistent image intensities across all subjects for the given MR channel by the following steps: (1) skull-stripping - using FSL [SJW04]; (2) B1 field correction and intensity normalization - using Freesurfer [SZE98] to standardize the intensity of MR images acquired from different medical centers.

In order to reduce the processing time, we applied the algorithms in a predefined volume of interest (VOI). For each 3D MR volume, the user visually identifies the start and end slice of the tumor, and provides manual seeds on the center slice of the tumor to initialize the GrowCut method. With this information, the VOI can then be generated. First, the bounding box of the input seeds on the tumor center slice is extended 25mm along each in-plane direction to enclose the whole tumor; then, the bounding box is extended in the z-direction to the start and end slice to obtain the VOI. Calculation time is thereby reduced by applying the segmentation framework only within the VOI instead of the whole brain volume.

We applied the proposed framework with the following parameter setup. For GrowCut, users were required to click 3-5 points in the foreground and background structures respectively on the center slice of the tumor. The FC algorithm used the same foreground seeds as GrowCut, and the $s1, m1, a1$, and $a2$ were chosen as described in the Method Section. Voxel classification using SVM did not utilize any user seeds. The SVM was trained for each leave-one-tumor-out iteration, resulting in 45 runs.

To obtain binary segmentation results, the outputs of the SVM, fuzzy connectedness, and the CMA ensemble were thresholded adaptively using the Otsu method [Ots79]. The binary segmentations from SVM and CMA were further processed by a connected component analysis to remove speckle noise. The post-processing analysis has two features: 1) the components smaller than 27 voxels are removed; 2) the components including background seeds but no foreground seeds are removed.

The accuracy of the segmentation result was evaluated by calculating the F_1-measure (ranging from 0-1) [Rij79] between the semi-automated segmentation result and the ground truth.

$$F_1{-}measure = 2{*}\frac{precision * recall}{precision + recall},\ with\ precision = \frac{tp}{tp + fp},\ recall = \frac{tp}{tp + fn}$$

3.4 Results

We calculated the F_1-measure for all 46 cases to evaluate the accuracy of the segmentation results against the ground truth, and to compare the performance of the three base methods and our ensemble method. We present the F_1-measure plot in Figure 3.2. Not a single base algorithm beats the other two algorithms in all 46 cases. GC performed best for 34 cases, while FC and SVM performed best for 7 and 5 cases, respectively. Our ensemble method was close to the best base result, even though the best base method varied for each case.

The ensemble method improved the F_1-measure by about 0.04 (0.04 ± 0.02) compared to the highest individual accuracy for eleven cases (No. 4, 8, 11, 12, 13, 14, 18, 36, 41, 43, 45), shown in Figure 3.2. Two main reasons for the improvement are observed. One is that when the tumor is inhomogeneously enhanced, the

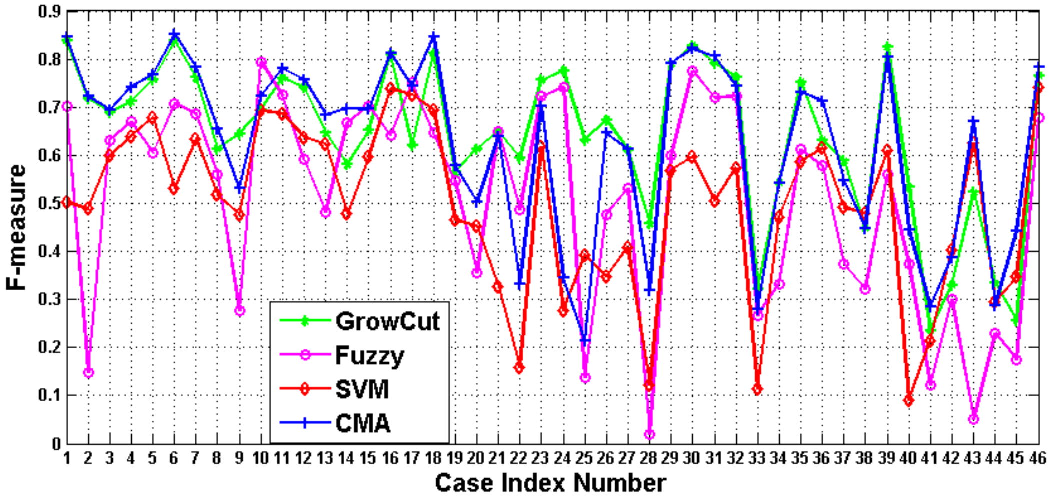


Figure 3.2: F_1-measure of three base methods and the ensemble method for all 46 cases

ensemble method included more tumor pieces than each base method. The other is that necrosis was often incorrectly included as tumor by GrowCut and FC method, but correctly removed by our ensemble method. Figure 3.6 shows an example (index No. 12).

The ensemble method performs similarly, improving the F_1-measure by 0.0006 ± 0.01 compared to the best performing method for twenty-one cases (No. 1, 2, 3, 5, 6, 7, 15, 16, 17, 19, 21, 27, 29, 30, 31, 32, 34, 35, 39, 42, 46), shown in Figure 3.2. We observed that one base method (GrowCut) performs relatively well for these cases, while the other two methods do not provide much additional value to our CMA method. In such cases, the tumor usually appears as a well-enhanced and single component, as shown in Figure 3.7 with index No. 31. Also, CMA selects more true positive voxels compared to the base methods, but also includes more false positive voxels. The cancelation leads to no overall improvement.

The ensemble method reduced the F_1-measure (-0.13 ± 0.14) of the individual methods in fourteen cases (No. 9, 10, 20, 22, 23, 24, 25, 26, 28, 33, 37, 38, 40, 44), shown in Figure 3.2. The reason is that the ensemble either missed more true

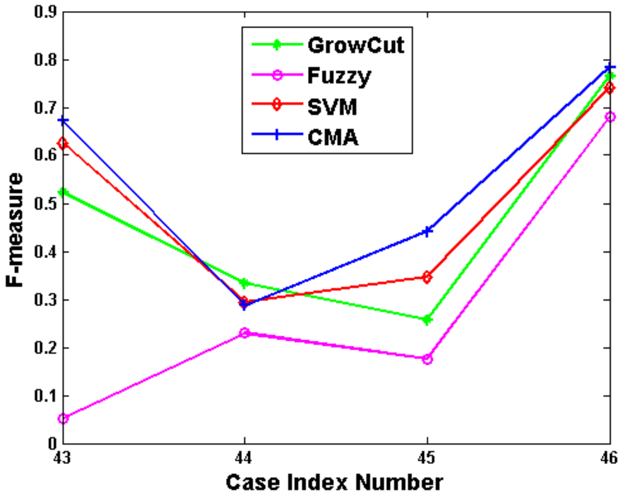


Figure 3.3: F_1-measure of three base methods and the ensemble method for the subgroup of multi-focal tumors

positive voxels, included more false positive voxels, or the number of false positive voxels exceeded true positive. For example, case No. 8 shown in Figure 3.8, had reduced the performance was reduced because many partial volume voxels are missed by our ensemble method.

Cases 43-46 in Figure 3.2 are multi-focal tumors, and are shown expanded in Figure 5.2. Multi-focal tumors are those with more than one lesion site, as defined by intervening areas of normal brain signal, including or excluding the primary site, all with a well-defined or mostly well-defined border [PXP11]. For this sub-group, GC missed unconnected tumor pieces and SVM included all tumor pieces. Our ensemble method improved the performance over the GC method by 0.08 ± 0.01, improved over the SVM method by 0.04 ± 0.04, and improved over the FC method by 0.26 ± 0.25. The ensemble method exhibited promising results that improve upon the best overall base method (GrowCut) in multi-focal tumors.

In general, the F_1-measure for all 46 cases is lower than 0.9, because the partial volumed voxels tend to be missed by the automated methods. Thus, the

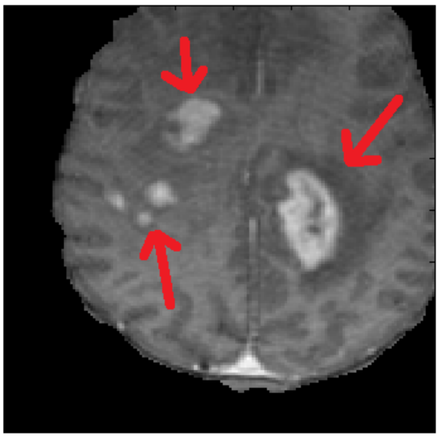


Figure 3.4: An example of multi-focal GBM tumor with red arrows pointing to the contrast enhanced multi-focal tumors

F_1-measure cannot exceed 0.9 even when the segmentation result is reasonably accurate by visual inspection.

The box plot of the the F_1-measure for all 46 cases is shown in Figure 3.5. The statistics of the F_1-measure over all 46 cases are summarized in Table 3.1 comparing base methods and the ensemble method. The median of the CMA method is slightly higher than that of all three base methods. We ran a paired t-test to compare the three base methods and the CMA ensemble method, and the results are shown in Table 3.2. It shows that GC, SVM and CMA are all significantly better than the FC method, and there is no significant difference between GC, SVM and CMA methods.

3.5 Discussion and Conclusion

In this study, we proposed an ensemble framework for GBM brain tumor segmentation on high-resolution T1w post contrast MR images. Instead of developing a customized method for this specific application, the proposed ensemble

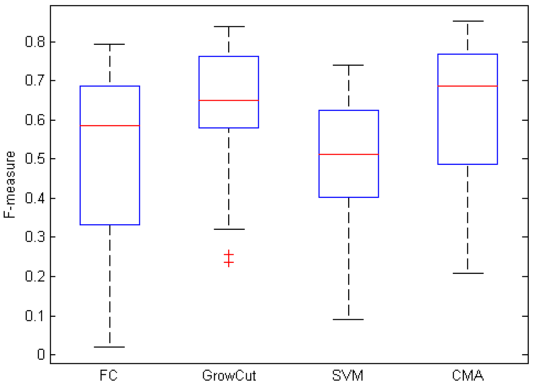


Figure 3.5: Box plot of F_1-measure for three base methods (FC, GrowCut and SVM) and the ensemble (CMA)

Table 3.1: Statistics of F_1-measure over 46 cases for different methods

Different methods	Mean	Median	STD	IQR
FC	0.51	0.59	0.22	0.35
GrowCut	0.64	0.65	0.16	0.18
SVM	0.50	0.51	0.17	0.22
CMA	0.63	0.7	0.18	0.27

method combines existing general-purpose segmentation algorithms to compensate for their respective advantages and weakness.

To our knowledge, we are the first to investigate ensemble segmentation for GBM tumor segmentation. In addition we evaluate the base methods and ensemble method on a dataset of 46 GBM tumors including different types of GBM tumor appearance patterns, which is substantially larger than reported 5-20 cases in the literature. It is necessary to evaluate the GBM tumor segmentation over a large dataset, because the appearance of GBM tumors on the images can vary substantially from case to case. GBM brain tumor segmentation is challenging

Table 3.2: Comparing different methods using paired t-test

	FC	GrowCut	SVM	CMA
FC	N/A	$p < 0.001$	$p < 0.05$	$p < 0.001$
GrowCut	$p < 0.001$	N/A	$p < 0.001$	$p > 0.05$
SVM	$p < 0.05$	$p < 0.001$	N/A	$p < 0.001$
CMA	$p < 0.001$	$p > 0.05$	$p < 0.001$	N/A

problem due to tumor heterogeneity, inhomogeneous intensity profiles, variable shapes and sizes, and different recurrent patterns after surgery. For example, there may or may not be necrosis/cavity/cyst present in the middle of the tumor; the tumor recurrence may occur in the primary site or at a distant site; the tumor may show a vivid enhancement or diffuse pattern; and the tumor could have a blob shape or irregular shapes. Thus, it is crucial to evaluate the segmentation method on a large dataset in a clinical setting. Liu et al. [LUO05] and Corso et al. [CSD08] are the only two studies in the literature that evaluated their systems on a dataset of 20 cases. In our study, we included 46 tumors, including cases with all the clinical conditions mentioned above. This study is thus significant in elucidating the range of tumor types to be addressed and thereby suggests that an ensemble approach may be appropriate.

To our knowledge, we are also the first to investigate ensemble segmentation on single-channel MR images (T1w post contrast) for GBM brain tumor segmentation. Most of previous studies developed fully-automated segmentation using multi-channel MR images (T1w, T2w, FLAIR, etc.), and we only find one publication, where Dube et al. [DCY08] did a preliminary study of automatic segmentation on this task using a dataset of 7 patients. In the setting of GBM tumor clinical trials, radiologists manually contour contrast-enhanced tumors on a single-channel T1w post contrast images to measure the tumor size change. Therefore, semi-automated segmentation on a single-channel T1w MR volume is relevant in a drug trial that uses radiographic response as a surrogate endpoint.

We compared the performance of different base segmentation algorithms on the application of GBM brain tumor segmentation. In the literature, many algorithms were proposed as general-purpose segmentation methods; however, it is hard to compare their performance since they were applied to different dataset. In this study, we evaluated three base algorithms on the same dataset, which serves as a reference to compare their performances and makes a useful contribution to the segmentation of GBM brain tumors.

Our study investigates a general-purpose segmentation framework, even though our ensemble method was tested for a specific application of GBM tumor segmentation. In the context of tumor drug clinical trials where radiographical response is used as a surrogate endpoint, imaging core labs need a general-purpose segmentation method for medical image segmentation, and the ensemble framework is a potential solution. This is because imaging core labs collect and process data from different trials with different diseases and image modalities (CT, MRI, PET, etc.). On one hand, it is tedious work for radiologists to manually contour the tumors; on the other hand, it is expensive and inefficient to design a specific segmentation algorithm for each application. Therefore, a general purpose segmentation framework becomes necessary in this context. However, medical image segmentation is not a trivial task due to the nature of medical image acquisitions and of heterogeneity of human diseases. An ensemble framework can take advantage of different and arbitrary segmentation algorithms. Its potential to serve as a general-purpose segmentation framework can be further studied and evaluated in other applications in the future to test the generalizability of the methods.

There are a couple of limitations in the present study design. One of them is that we only had one radiologist's reading as the ground truth. In the future, we will have multiple reader markings as ground truth, and compare the performance of the ensemble framework in terms of inter-reader reproducibility. The other limitation is that the number of input seeds for FC and GrowCut methods were not strictly controlled among all cases. For future work, we will design a tightly-controlled rule for the input seeds to compare the two methods, and vary the number of seeds to study their repeatability.

Future work could also be done to improve our CMA ensemble method. One possibility is to assign different weights to each base segmentation algorithm. Currently, all algorithms are equally weighted. In the future, we will explore ways to associate different confidence coefficients to each base method. Another possibility is to include additional base methods to explore whether more base methods can improve the segmentation performance.

In summary, we compared three base segmentation methods and our ensemble method on a GBM dataset of 46 cases, and found that ensemble segmentation does not necessarily improve segmentation accuracy upon base methods if the base methods have similar advantages and are not complementary to each other. In such a case, we found that the ensemble segmentation result is very similar to the best performed method (GrowCut). This provided motivation to make ensemble segmentation a future potential solution for GBM tumors which have a variety of appearances. With properly selected base methods which are good at segmenting different type of tumor appearances, an appropriate ensemble method may sustain the accuracy from the best "performer" for different tumor appearance and achieve an overall improvement over the base methods.

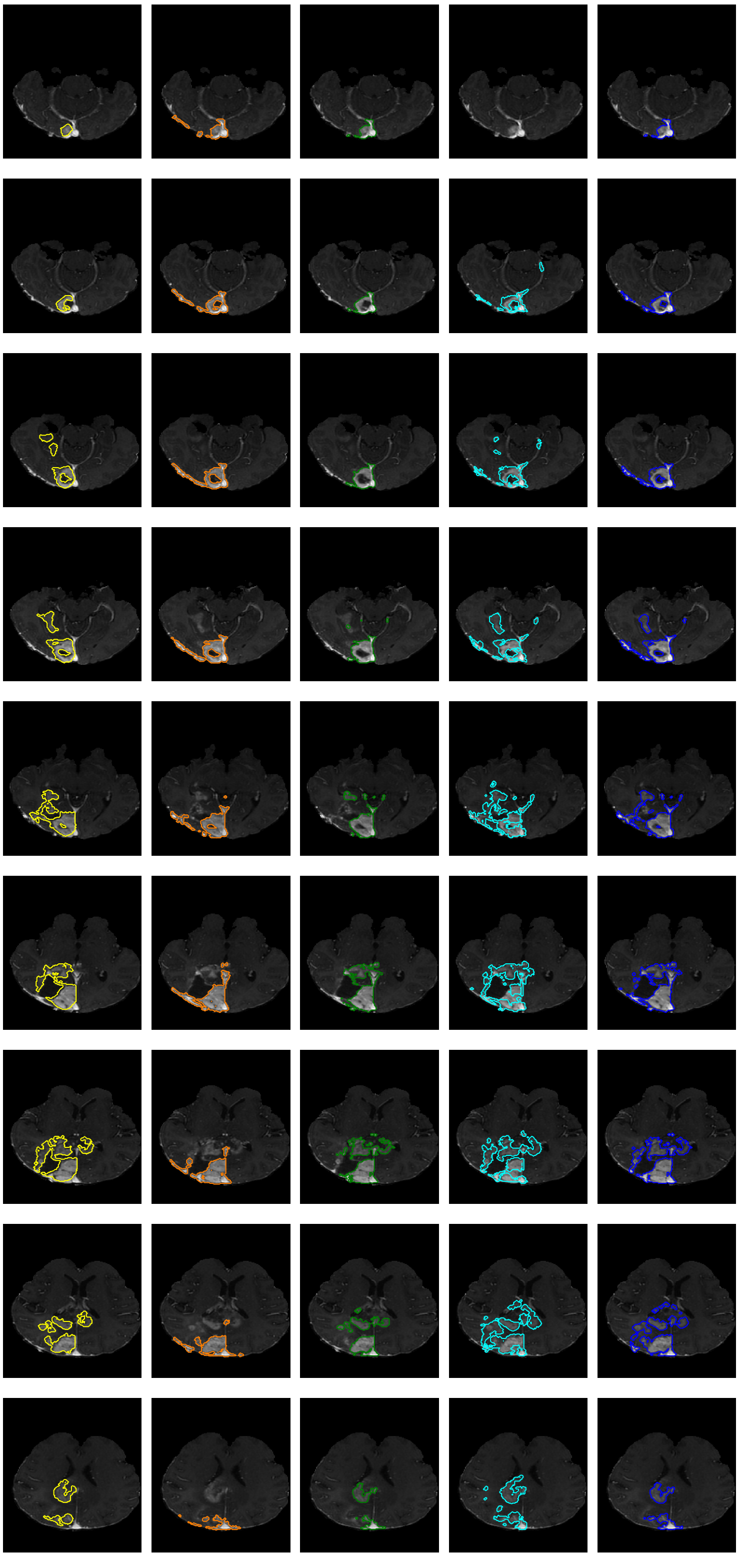


Figure 3.6: Illustrative example of the segmentation results of the tumor with index number 12. Rows show results on different slices; columns show results using different segmentation methods: yellow - the ground truth; orange - FC; dark green - GC; light blue - SVM; dark blue - CMA ensemble

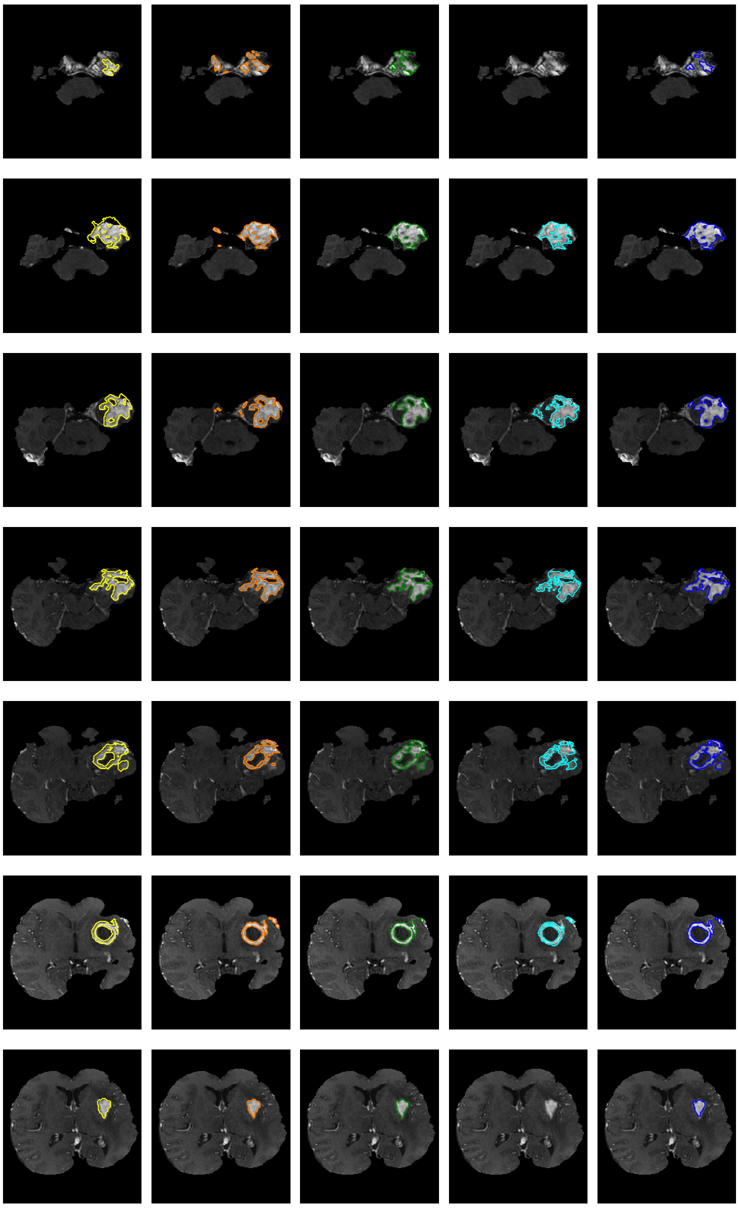


Figure 3.7: Illustrative example of the segmentation results of the tumor with index number 31. Rows show results on different slices; columns show results using different segmentation methods: yellow - the ground truth; orange - FC; dark green - GC; light blue - SVM; dark blue - CMA ensemble

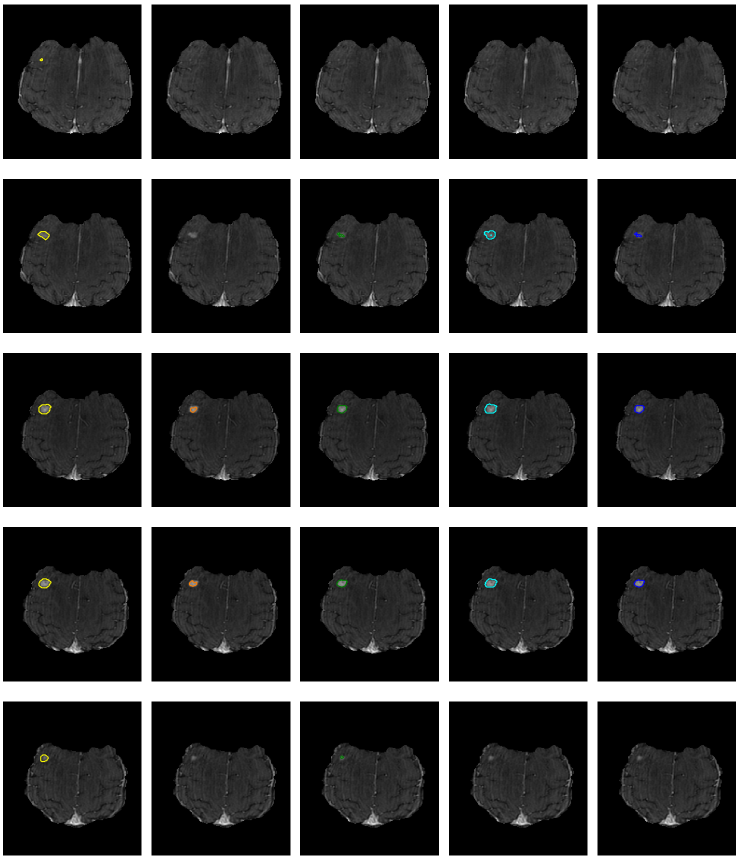


Figure 3.8: Illustrative example of the segmentation results of the tumor with index number 8. Rows show results on different slices; columns show results using different segmentation methods: yellow - the ground truth; orange - FC; dark green - GC; light blue - SVM; dark blue - CMA ensemble

CHAPTER 4

Improving Semi-automated Segmentation by Integrating Learning with Focused Sampling

Abstract

Interactive segmentation algorithms such as GrowCut usually require numerous user interactions to perform well for complex tumors, and have poor reproducibility with different initialization. In this study, we developed a technique to boost the performance of the interactive segmentation method involving: 1) a novel focused sampling scheme for supervised learning, as opposed to conventional random sampling; 2) boosting GrowCut using the machine learned results instead of additional manual inputs. We applied the proposed technique to glioblastoma multiforme (GBM) brain tumor segmentation, and evaluated the technique on a preliminary dataset of ten randomly-selected cases from a multi-center pharmaceutical drug trial. The results showed that the proposed system has the potential to reduce user interaction while maintaining similar segmentation accuracy, and improved reproducibility.

4.1 Introduction

Two weaknesses common to interactive segmentation algorithms include the need for excessive user interactions and a lack of repeatability. The lack of repeatability is due to user interaction variations. Excessive user interaction means that a large number of user inputs are required to obtain satisfactory results, reducing the usability. To overcome these problems, we developed a method to boost an existing interactive segmentation algorithm (specifically GrowCut [VK05]) with machine learning results using focused sampling, and applied the proposed method to glioblastoma multiforme (GBM) brain tumor segmentation on T1w post-contrast images.

GBM brain tumors usually consist of three parts: active tumor, necrosis and edema. In this study, the goal is to segment the active tumor part, which is contrast-enhancing component on T1w post-contrast MR images, as used in the RANO criteria [WMR10] for treatment response evaluation. The goal is the same as in Chapter 2 and Chapter 3.

4.1.1 Limitations of GrowCut on GBM Segmentation

The GrowCut method [VK05] is an interactive segmentation technique using the cellular automata theory, and it is explained in details in Chapter 3.2.1.2. There are several problems when applying GrowCut to GBM tumor segmentation as shown in Figure 4.1. With limited user input seeds on an active tumor component, first, the necrosis as the dark core is incorrectly labeled; second, for multifocal tumors where there are unconnected tumor pieces, GrowCut cannot detect all of them. Furthermore, on post-contrast MR images, both active tumors and non-target vessels and dura have similar bright intensity, thus, GrowCut fails to differentiate between them, as shown in Figure 4.1. So for the GrowCut method, users often need to manually place numerous additional seeds to update the GrowCut segmentation results and generate accurate results.

4.1.2 Literature

There has been little investigation into improving interactive segmentation with machine learning methods. Miller et al. [VM11] used active learning to reduce

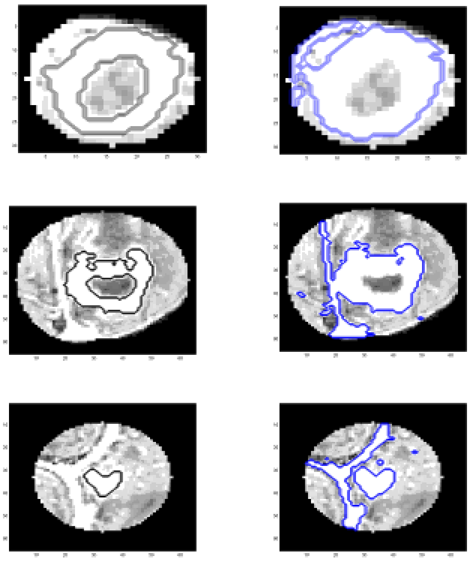


Figure 4.1: Limitations of GrowCut algorithm: left - ground truth segmentation; right - GrowCut segmentation result with a single seed

the GrowCut user interaction. This approach automatically suggest placement of user interactions for the user to review. Top et al. [VM11] incorporated active learning in order to assist the user in choosing where to provide interactive input by automatically suggesting the the plane of maximal uncertainty. Both systems focus on suggesting more informative voxels or planes for user to provide input, while our system focuses on automatically generating seeds to eliminate the need for additional user input.

4.1.3 Contributions

In this study, we developed a novel method to boost the original GrowCut method. The goal is to reduce the amount of user input and improve the segmentation reproducibility. First, we applied machine learning methods to provide additional automatic seeding. Second, in building the machine-learned classifier, we developed a focused sampling scheme, in contrast to conventional random sampling.

Focused sampling involves collecting a larger number of difficult training voxels and a smaller number of easy training voxels. The hypothesis is that by integrating machine learning with focused sampling, the boosted GrowCut will reduce the user interaction while maintaining similar accuracy, and improve reproducibility. To our knowledge, ours is the first to employ focused sampling to aid interactive segmentation for medical images.

4.2 Methods: Boosting GrowCut with Learning Results from Focused Sampling

The brain volume was first preprocessed to remove non-brain matter and obtain consistent image intensities across all subjects for the given MR channel by the following steps: (1) skull-stripping - using FSL [SJW04]; (2) B1 field correction and intensity normalization - using Freesurfer [SZE98] to standardize the intensity of MR images acquired from different medical centers. The pre-processing pipeline is the same as in chapter 2 and chapter 3.

4.2.1 Supervised Learning with Focused Sampling

4.2.1.1 Training Set Collection by Focused Sampling

We define focused sampling as focusing on difficult sub-classes during the training sample collection. In conventional random sampling, a certain number of tumor voxels are randomly sampled within the ground truth tumor region and equal number of voxels are randomly sampled from the rest of the brain tissue. The background class includes sub-classes of brain white matter, gray matter, dura, vessel, cerebrospinal fluid (CSF), etc. Within these subclasses, CSF, white matter and gray matter are relatively easy to be differentiated from tumors, while dura and vessels are difficult. Sufficient training samples from difficult subclasses are

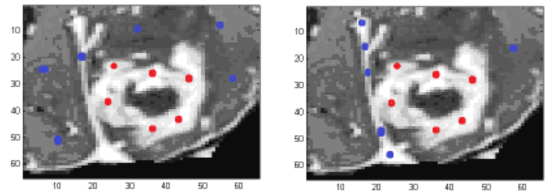


Figure 4.2: Example of difference between the random sampling (left) and focused sampling (right).

necessary to build an effective classifier. However, Conventional random sampling cannot guarantee enough training samples from each subclass. Thus, we developed the focused sampling scheme, by which we intentionally build a training set with more weight given to difficult sub-classes and less weight to easy sub-classes. This involves collecting more training voxels for difficult sub-classes, and less training voxels for easy sub-classes. Figure 4.2 shows one example of the difference between random sampling and focused sampling.

The idea of focused sampling is inspired by active learning. In active learning, the machine learned classifier automatically picks the difficult samples and requires the oracle (eg, a human annotator) to provide annotation; while in our focused sampling scheme, readers inherently collect difficult samples for the machine learner, while using interactive segmentation method to be boosted.

4.2.1.2 Collecting Focused Samples

Focused sample collection is achieved naturally by users of the original GrowCut method. Users need to follow four steps to use the GrowCut method: 1) Manually click one seed on an active tumor part and one seed on the background; 2) Run the GrowCut method; 3) Review the segmentation result, and successively place more seeds on the false-segmented structures or seeds on missed tumor components, rerun GrowCut and update the result; 4) Repeat step 3 until the user finds nothing more to modify. Figure 4.3 shows one example of using the original GrowCut method and providing additional seeds to achieve final accurate results.

For focused sampling, all the seeds user clicked and painted during the procedure are collected as the training set. As described in step 3), the user needs to manually paint additional seeds either on the false-segmented structures or on missing components to correct the segmentation results. Thus, these additional seeds are from the difficult structures that original GrowCut could not handle and thus naturally become focused samples.

The advantage of using collected focused samples for training is that they are the most representative samples of the difficult structures for GrowCut. Instead of sampling from the entire brain, we focus on the structures which are most difficult for the GrowCut method. Therefore, the proposed classification system will not do a simple tumor versus non-tumor classification, instead, it is the classification of tumor vs difficult non-tumor structures.

4.2.1.3 Feature Calculation

For each training sample, a set of features were calculated: intensity, gradient magnitude, first-order Gaussian derivatives (in three directions), second-order Gaussian derivatives (six in total), and the three eigenvalues of the Hessian matrix on scales 1, 2, and 4, and derived features using the three eigenvalues, resulting in 73 features in total.

The three eigenvalues of the Hessian matrix are represented as λ_1, λ_2, and λ_3, and are ranked according to their absolute value $|\lambda_1| > |\lambda_2| > |\lambda_3|$. The derived features based on the three eigenvalues are calculated to describe the local shape [OGA07]: the magnitude of the eigenvalues, $\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}$, and the ratios between the eigenvalues (λ_2/λ_1, λ_3/λ_1, $\frac{|\lambda_1|-|\lambda_2|}{|\lambda_1|+|\lambda_2|}$, and $\frac{|\lambda_3|}{\sqrt{|\lambda_1\lambda_2|}}$).

Multiscale analysis is performed by computing the Hessian matrix (and eigen-

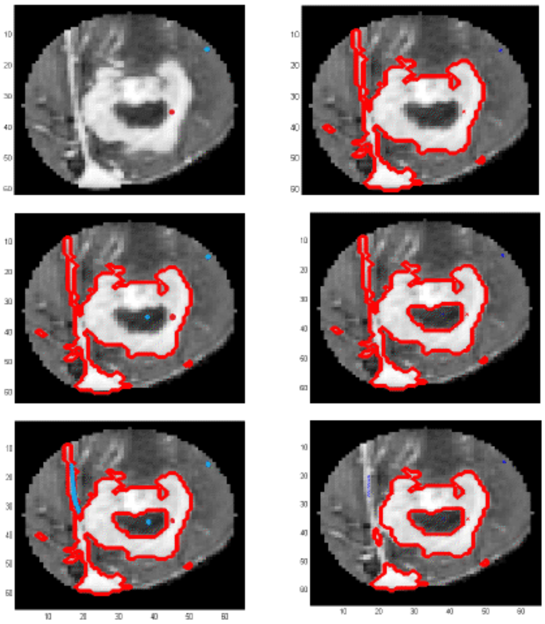


Figure 4.3: Illustration of providing additional seeds required for accurate segmentation results using GrowCut method. The red dots are tumor seeds and the blue dots are background seeds. The red contour is the segmentation result. The left column shows the user inputs at each step, and the right column shows the segmentation result based on the user input on the left.

values) at multiple scales. The scale σ is the standard deviation of the Gaussian smoothing kernel. In this study, $\sigma = 1, 2, 4$ *pixels*. The range of scales allows the shape of both larger and smaller objects to be detected and quantified. At lower scales, the shape of larger structures may not be accurately captured due to noise and small inhomogeneities in the structure. At higher scales, the shape of smaller objects may be distorted as neighboring structures are smoothed together. The shape features are computed at each scale along with the max and min (based on magnitude) of the feature over all scales. In total, we have 73 features calculated for each voxel.

4.2.1.4 Classifier

We ran a leave-one-tumor-out cross validation for the classifier. In each run, we applied principle component analysis (PCA) [DHS00] to reduce the feature dimensionality, and train a linear discriminate classifier (LDC) [DHS00] to differentiate tumor voxels from non-tumor voxels. We used the LDC classifier implementation in prtools [Dui00].

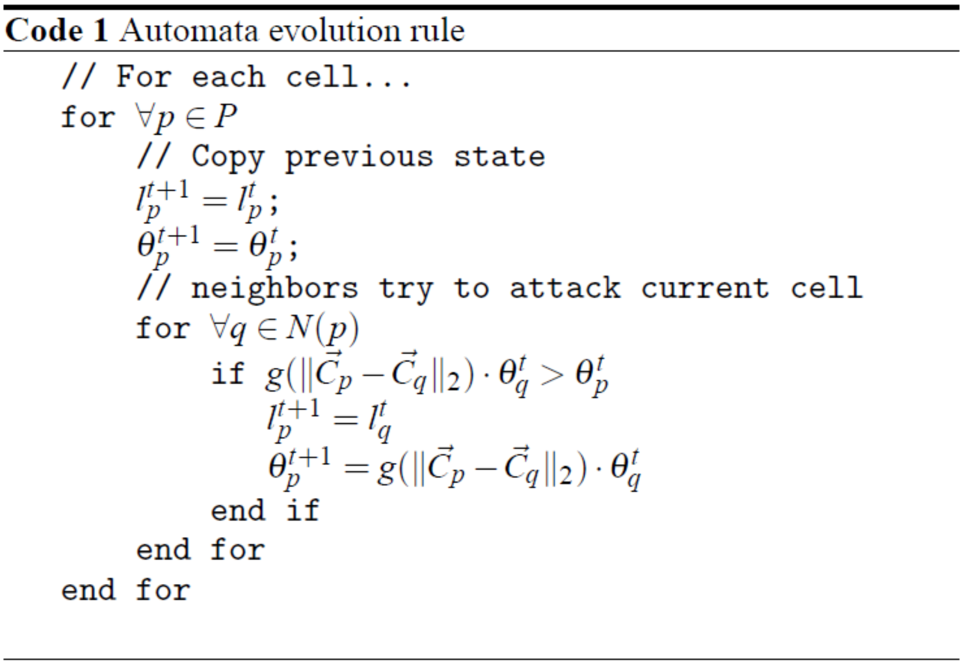
4.2.2 Boosting GrowCut with Learning Results from Focused Sampling

4.2.2.1 Original GrowCut

The GrowCut method [VK05] is an interactive segmentation technique using the cellular automata theory, and it is explained in detail in Chapter 3.2.1.2. The method starts from user-clicked seeds on both object and background. For each voxel, it allows the neighboring voxels to attack, and the strength of attacking is based on the local neighbor similarity. The neighbor with the strongest attacking strength will update the current voxel with the same label. The pseudo code in each iteration is shown in Figure 4.4. The method converges when the label map and strength map do not change any more. The method assigns each voxel both a label and a strength value. The label indicates the class, while the strength shows how confident it is about the labeling.

4.2.2.2 Boosting GrowCut Algorithm

Given a new GBM tumor case, there are two phases to run the boosting GrowCut algorithm: first, apply the machine learned LDC classifier on the new test case; second, the reader paints initial seeds and runs the boosting GrowCut method as described in Algorithm 1 of Figure 4.5. We described our algorithm using the



Code 1 Automata evolution rule

```
// For each cell...
for $\forall p \in P$
    // Copy previous state
    $l_p^{t+1} = l_p^t$;
    $\theta_p^{t+1} = \theta_p^t$;
    // neighbors try to attack current cell
    for $\forall q \in N(p)$
        if $g(\|\vec{C}_p - \vec{C}_q\|_2) \cdot \theta_q^t > \theta_p^t$
            $l_p^{t+1} = l_q^t$
            $\theta_p^{t+1} = g(\|\vec{C}_p - \vec{C}_q\|_2) \cdot \theta_q^t$
        end if
    end for
end for
```

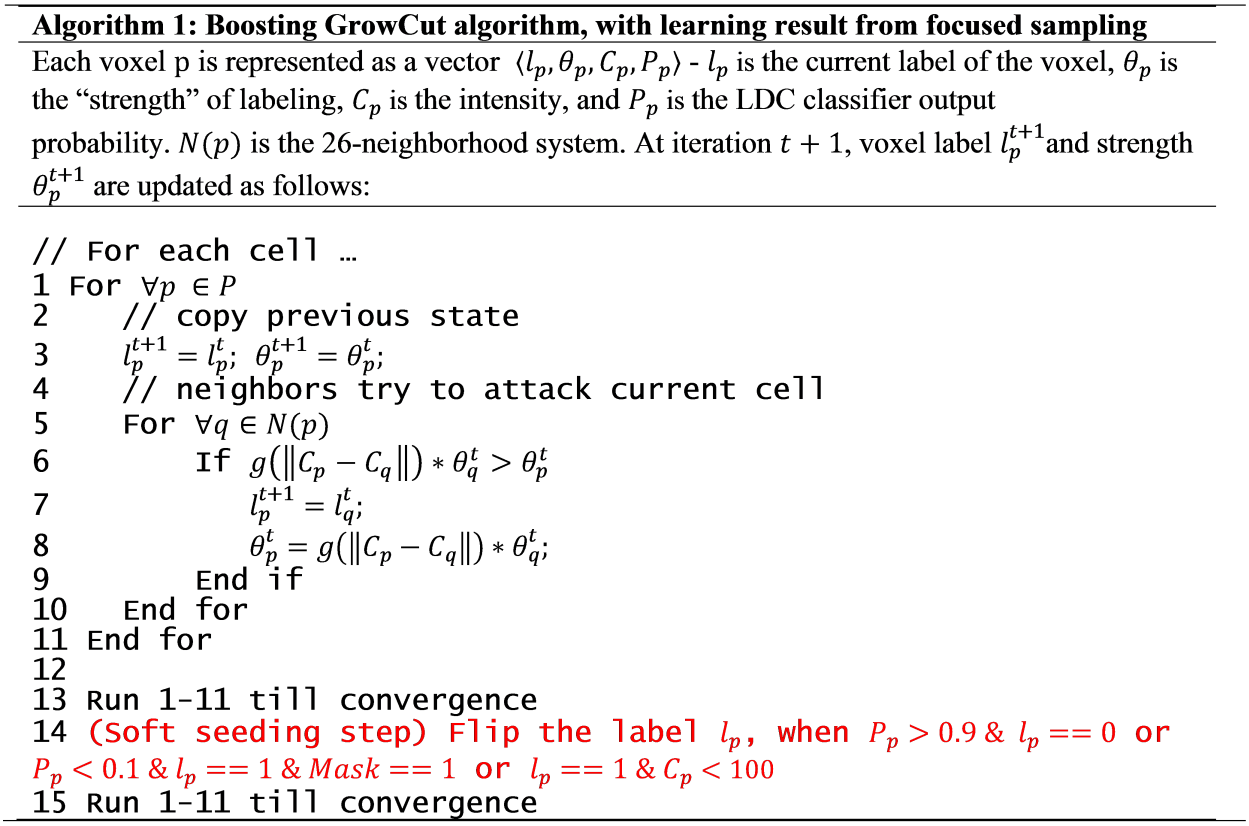
Figure 4.4: Original GrowCut algorithm - adapted from Vezhnevets et al. [VK05]

same format as in the original GrowCut paper for comparison, and our boosting steps are highlighted in red, where $g(x)$ is a monotonically decreasing function bounded to $[0, 1]$.

4.2.2.3 Automatic Soft Seeding

We use the machine learned probability to automatically generate seeds to refine the segmentation results instead of manual input seeds. Soft seeding means that we set the strength of the automatic seeds to be the machine learned probability, in contrast to setting the strength of the manual input seeds to be 1.0. Line 14 in Algorithm 1 defines the rules for automatic soft seeding based on the machine learned results. The details are explained as follows.

$P_p > 0.9$ and $l_p == 0$: Voxels with high probability of being tumor are selected as automatic tumor seeds. For these voxels, update the label map to be tumor,



Algorithm 1: Boosting GrowCut algorithm, with learning result from focused sampling

Each voxel p is represented as a vector $\langle l_p, \theta_p, C_p, P_p \rangle$ - l_p is the current label of the voxel, θ_p is the “strength” of labeling, C_p is the intensity, and P_p is the LDC classifier output probability. $N(p)$ is the 26-neighborhood system. At iteration $t+1$, voxel label $l_p^{t+1}$and strength θ_p^{t+1} are updated as follows:

```
// For each cell …
1 For $\forall p \in P$
2    // copy previous state
3    $l_p^{t+1} = l_p^t;\ \theta_p^{t+1} = \theta_p^t;$
4    // neighbors try to attack current cell
5    For $\forall q \in N(p)$
6        If $g(\|C_p - C_q\|) * \theta_q^t > \theta_p^t$
7            $l_p^{t+1} = l_q^t;$
8            $\theta_p^t = g(\|C_p - C_q\|) * \theta_q^t;$
9        End if
10   End for
11 End for
12
13 Run 1-11 till convergence
14 (Soft seeding step) Flip the label $l_p$, when $P_p > 0.9$ & $l_p == 0$ or
$P_p < 0.1$ & $l_p == 1$ & $Mask == 1$ or $l_p == 1$ & $C_p < 100$
15 Run 1-11 till convergence
```

Figure 4.5: Boosting GrowCut algorithm

and update the strength map to be the LDC probability. The goal is to add unconnected tumor components which are easily missed by the original GrowCut.

$P_p < 0.1$ and $l_p == 1$ and $mask == 1$: Voxels with low probability of being tumor within a mask region are selected as automatic background seeds. For these voxels, we update the label map to be non-tumor and update the strength map to be $(\ 1 - LDC\ probability)$. The mask is the automatically detected dural region, which will be explained in the next section. The goal is to remove dural or vessel structures.

$l_p == 1$ and $C_p < 100$: Voxels with intensity lower than brain white matter are selected as automatic background seeds. The goal is to remove necrotic regions.

4.2.2.4 Automatic Detection of Dural Region Mask Using Hough Transform

The dura appears to have high image intensity after contrast injection and is indistinguishable from tumors by the original GrowCut method. We automatically detect dural regions to limit the search region for automatic soft seeding.

The dura is the outermost of the three layers of the meninges surrounding the brain and spinal cord. The dura surrounding the brain and the spinal cord is responsible for keeping in the cerebrospinal fluid [DVM09]. In this study, we focus on the dural regions either in between the two cerebral hemispheres (also known as falx cerebri) or covering the surface of the brain. For the dura located in the longitudinal cerebral fissure between the hemispheres, we run the Hough transform to detect it. For the dura covering the brain surface, we erode the skull stripping results, and the eroded region is saved as the dural region.

The Hough transform is the technique used to detect a certain shape in an image. In this study, we use it for line detection. Consider a point (x_i, y_i) and the general equation of a straight line in slope-intercept form, $y_i = ax_i + b$. However, writing this equation as $b = -x_i * a + y_i$ and considering the parameter space yields the equation of a single line for a fixed pair (x_i, y_i). Furthermore, a second point (x_j, y_j) also has a line in parameter space associated with it, and this line intersects the line associated with (x_i, y_i) at (a, b), where a is the slope and b is the intercept of the line containing both (x_i, y_i) and (x_j, y_j) in the xy plane. In fact, all points contained on this line have lines in parameter space that intersect at (a, b). A problem with using equation $y = a * x + b$ to represent a line is that both the slope and intercept approach infinity as the line approaches the vertical. Thus, one can use the normal representation of a line: $xcos(\theta)+ysin(\theta) = \rho$. Then subdivide the parameter space into so-called accumulator cells, where $(\theta_{min}, \theta_{max})$ and (ρ_{min}, ρ_{max}) are the expected range of slope and intercept values. The cell at coordinate (i,j), with accumulator value $A(i,j)$, corresponds to the square associated with parameter space coordinates (θ_i, ρ_i). Initially, these cells are set to zero. Then, for every point (x_k, y_k) in the image plane, we let the parameter θ equal each of the allowed subdivision values on the theta axis and solve for the corresponding ρ using the equation $\rho = x_k * cos(\theta) + y_k * sin(\theta)$. The resulting ρ is then rounded off to the nearest allowed value in the ρ axis. If a choice of θ_i results in solution ρ_i, we let $A(i,j) = A(i,j) + 1$. At the end of the procedure, a value of M in $A(i,j)$ corresponds to M points in the xy plane lying on the line $\rho = xk * cos(\theta) + yk * sin(\theta)$. The accuracy of the collinearity of these points is determined by the number of subdivisions in the $\theta\rho$ plane.

There are two steps to detect the centerline of the brain using the Hough transform. First, 2D centerlines are detected slice by slice for certain slices of images located in the upper half of the brain. These slices are currently manually selected for each brain volume, and will be automated in the future using physical locations of the image slices. Then the points on the detected lines are fitted by a 3D plane using the least mean square method. The 2D line detection on one slice is performed by the following steps: 1) The 2D image is rotated by 30 degrees for better detection because the centerline is originally almost vertical. 2) The Canny edge detector is run on the 2D image and edges are detected and output as a binary mask. 3) Next, the Hough transform is applied to detect lines. 4) Post processing is applied to remove those lines that are not center lines. The rules are: the slope must fall into the range [1,2]; after rotating the line to strictly vertical, the x-coordinate has to fall into the range 120 ± 8. Lines that do not obey to the rules are discarded. The 3D points on the detected lines are collected, and 3D plane is fitted to the 3D point cloud using the least mean square technique.

For the dataset of 10 cases, 4 cases have excellent results by visual inspection, 5 have good results, and one fails because the dura is not well enhanced. The failing case is not a problem, since dura is not enhanced and thus is distinguishable from

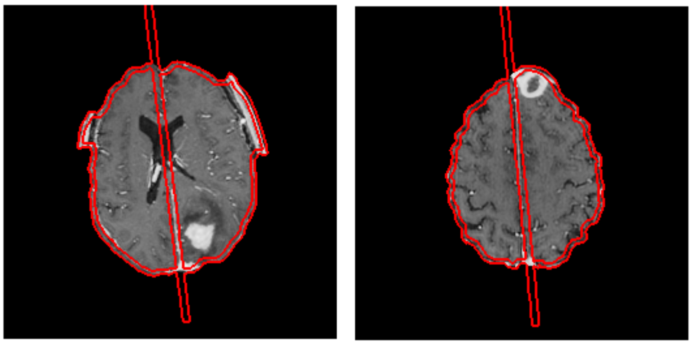


Figure 4.6: Examples of automatically detected dural regions

tumors. Figure 4.6 shows two examples of detected dural regions.

4.3 Experiments and Results

We have 10 GBM tumor cases from three medical centers in this study, randomly selected from our in-house research database. The ground truth for the segmentation is manually contoured by a board-certified neuroradiologist with 10 years of experience, with the facilitation of a semi-automated segmentation tool [Ots79] from an in-house software system QIWS (Quantitative Imaging Work Station).

The imaging protocol for the T1w sequence is 3D volumetric acquisition in the axial plane with flip angle-spoiled gradient echo sequence (FSPGR) or magnetization-prepared rapid gradient-echo (MP-RAGE) sequence with 1mm slice thickness, 0.9mm by 0.9mm pixel size, and 256*256 in-plane resolution.

The brain volume is preprocessed to remove non-brain matter and obtain consistent image intensities across all subjects for the given MR channel by the following steps: (1) skull-stripping - using FSL [SJW04]; (2) B1 field correction and intensity normalization - using Freesurfer [SZE98] to standardize the intensity of MR images acquired from different medical centers. The pre-processing is the same as described in Chapter 3.

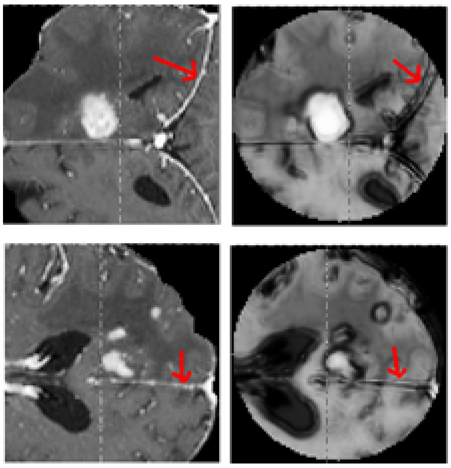


Figure 4.7: Learning results with focused sampling. Rows: examples of tumors; Columns: original MR image slice and the classifier probability output. Red arrows show the dura.

In order to reduce the processing time, we apply the algorithms in a pre-defined volume of interest (VOI). The protocol for determining the VOI is described in the Appendix.

We run a leave-one-tumor-out cross validation, that is, we randomly select 9 cases as training set, and use the rest one as a test case. During the training phase, for each of the 9 training cases, 2 voxels from each stroke of a user's paintings are randomly sampled to add into the focused training set. For the one test case, we extract the same imaging features for each voxel, input the feature vector into the trained classifier, and then obtain the posterior probability of being a tumor voxel.

Figure 4.7 illustrates the LDC classification result using focused sampling. The images on the left column show two examples of T1w post contrast MR image with red arrows pointing at the non-target brain structure (dura). The images on the right show the relevant classifier outputs from focused sampling - the probability of labeling a particular voxel as tumor. The dura has similar contrast enhancement to tumors and thus can not be correctly segmented by the original GrowCut, and voxel classification with random sampling cannot differentiate dura from tumor because there are not enough training samples from the dura. On the other hand, voxel classification with focused sampling can distinguish tumors from the dura, by assigning tumor high probability and dura low probability, because of the local shape features and the sufficient training samples from the dura structure by focused sampling.

After we obtain the posterior probability, we incorporate it into the boosting GrowCut method. We evaluate the boosting GrowCut method by comparing the proposed boosting GrowCut and the original GrowCut. We gradually add user seeds (two seeds at a time) and plot the segmentation accuracy versus seed number. The accuracy of the segmentation is evaluated by calculating the overlap ratio (as defined in Chapter 2 as $OR = \frac{A \cap B}{A \cup B}$) between the ground truth and the segmentation results, with A and B as binary segmentation results.

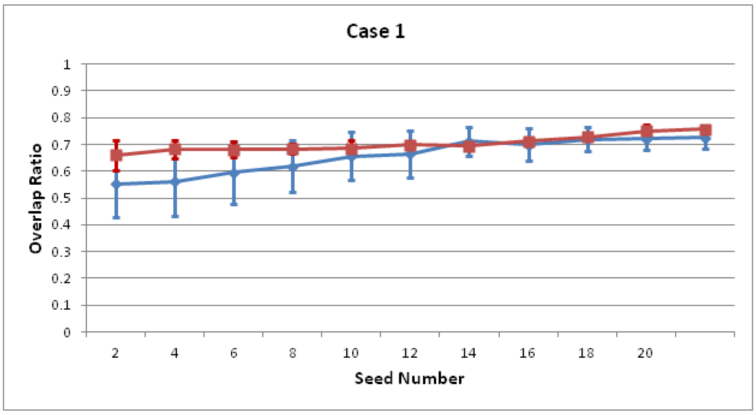
The original GrowCut and boosting GrowCut are evaluated independently using the following experiment. First, the segmentation is initialized by one tumor seed and one non-tumor seed. Original GrowCut (or boosting GrowCut) is applied respectively. Second, one additional seed for tumor and one for non-tumor is added to update the segmentation result using original GrowCut (or boosting GrowCut). Then, additional seeds are contiguously added in the same fashion for ten iterations, and the segmentation accuracy in each iteration is plotted for original GrowCut and boosting GrowCut as shown in Figure 4.8, where the y-axis is the overlap ratio, and the x-axis is the number of seeds, with blue plots as the original GrowCut and red plots as the boosting GrowCut. The experiment is repeated 20 times to generate the error bars on the plot in Figure 4.8. The average overlap ratio of the 10 cases is presented in Figure 4.9, and paired t-test shows there that boosting GrowCut method is significantly better than the original GrowCut ($p < 0.001$). Examples of the segmentation results of cases 3, 4, and 1 are illustrated in Figure 4.3.

The ideal way of obtaining seeds would be a user placing manual additional seeds in each of the 10 iterations, however, in this preliminary study, we let the computer generate simulated seeds using the ground truth tumor contours. The initial simulated seeds are randomly picked from tumor and non-tumor regions, excluding the “band region” which is defined as the region between the 3-voxel erosion and 3-voxel dilation of the tumor contour, in order to avoid partial volume voxels. The additional simulated seeds are randomly picked according to the difference between the result of segmentation and the ground truth excluding the “band region”.

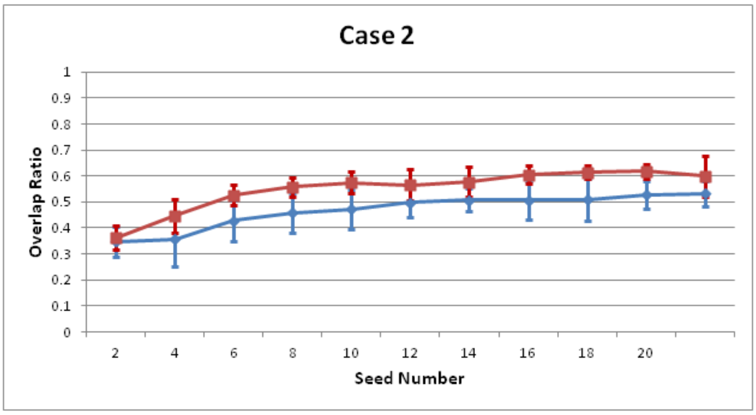
4.4 Discussion

In this study, we developed a framework for boosting GrowCut, which incorporates machine learning results into the original GrowCut method to reduce the amount of manual user interaction and improve the algorithm consistency. First of all, a machine learned classifier is trained using focused samples, in contrast to conventional random sampling. Second, for a new brain tumor, we apply the trained classifier to the new case to obtain the posterior probability. The boosting GrowCut is still used as interactive segmentation method and initialized by user input seeds. It incorporates machine learning results to generate automatic additional seeds. We evaluated the boosted GrowCut method to segment GBM tumors on T1w post contrast MR images.

Semi-automated segmentation may be useful for GBM brain tumor in clinical practice. There are quite a few full-automated methods but the accuracy is not satisfactory since GBM tumor has a large variation in the appearance and shape. The contrast enhancement is quite heterogeneous within the same tumor. These factors make GBM tumor segmentation a very challenging task and make it hard for a fully automated method to achieve good accuracy. Manual contouring is a

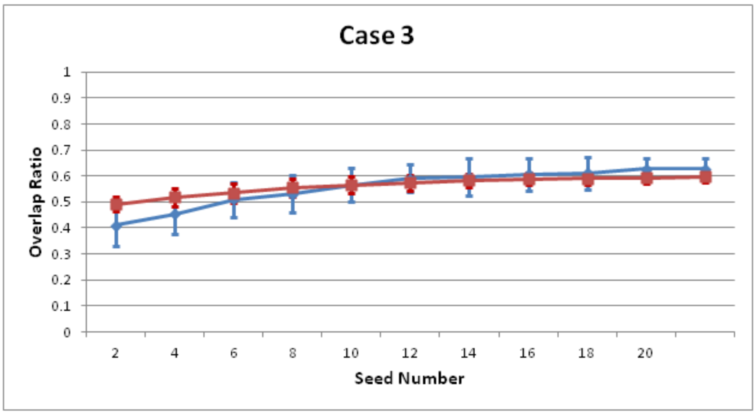


(a)

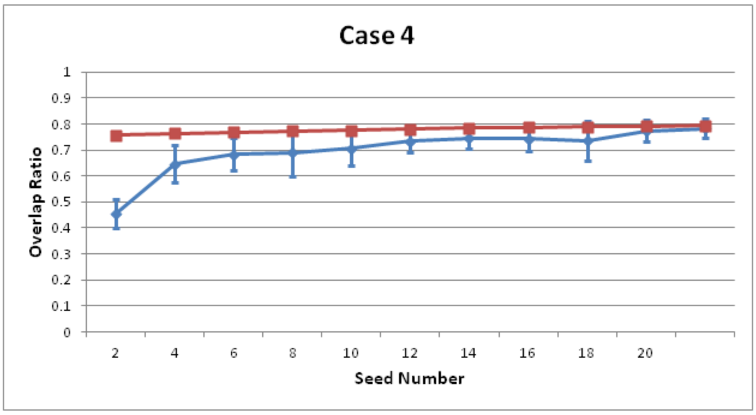


(b)

Figure 4.8: Comparing the original GrowCut (blue dots) and the boosting GrowCut (red squares). At each iteration, one tumor and one background seed is added to improve the segmentation result. The error bars are generated by 20 runs of random initialization. (continued)

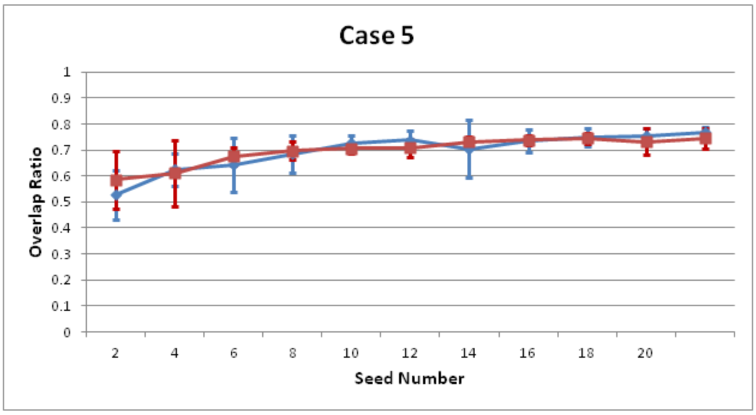


(c)

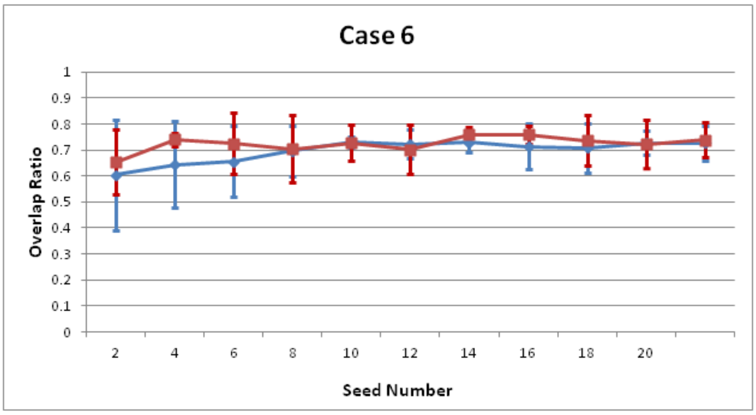


(d)

Figure 4.8: (continued) Comparing the original GrowCut (blue dots) and the boosting GrowCut (red squares). At each iteration, one tumor and one background seed is added to improve the segmentation result. The error bars are generated by 20 runs of random initialization.

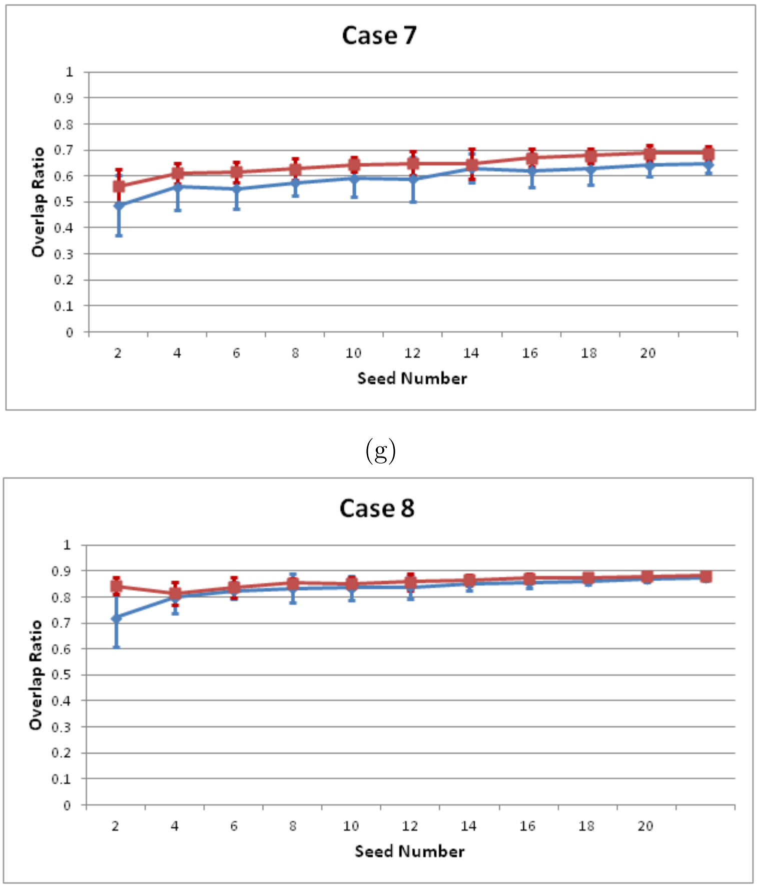


(e)



(f)

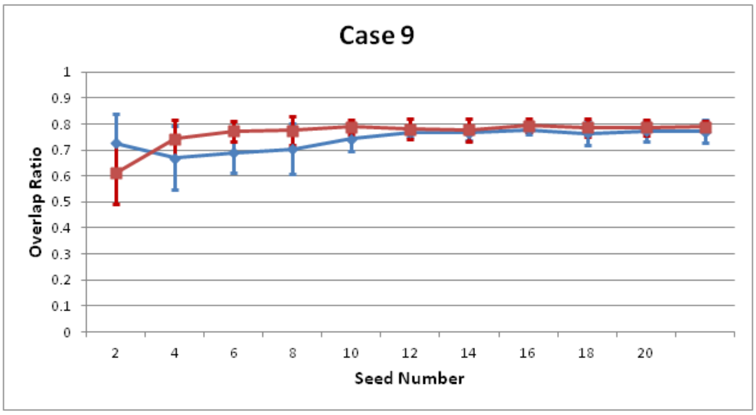
Figure 4.8: (continued) Comparing the original GrowCut (blue dots) and the boosting GrowCut (red squares). At each iteration, one tumor and one background seed is added to improve the segmentation result. The error bars are generated by 20 runs of random initialization.



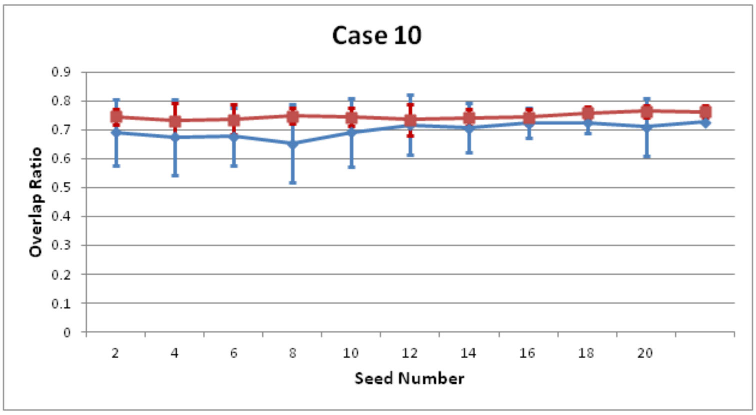
(g)

(h)

Figure 4.8: (continued) Comparing the original GrowCut (blue dots) and the boosting GrowCut (red squares). At each iteration, one tumor and one background seed is added to improve the segmentation result. The error bars are generated by 20 runs of random initialization.



(i)



(j)

Figure 4.8: (continued) Comparing the original GrowCut (blue dots) and the boosting GrowCut (red squares). At each iteration, one tumor and one background seed is added to improve the segmentation result. The error bars are generated by 20 runs of random initialization.

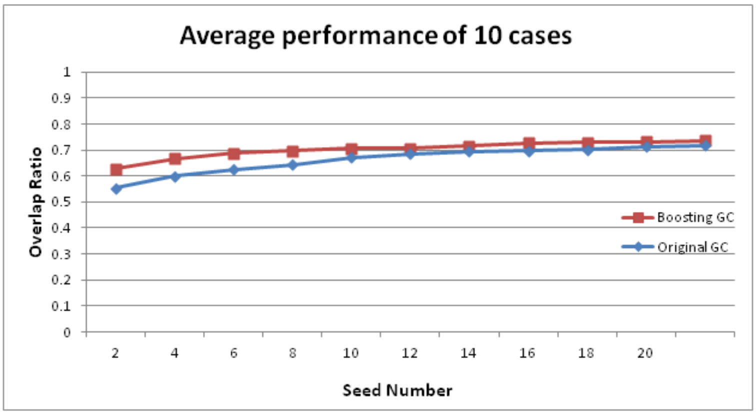


Figure 4.9: The average overlap ratio of 10 cases

very tedious task. Thus, we develop this semi-automated framework to allow for user interaction to edit the segmentation, and add machine intelligence to further reduce the user interaction.

For cases 1 and 2, boosting GrowCut improves the accuracy over the original GrowCut because the tumor is attached to the dura mater on the brain surface, which is very hard to distinguish using intensity for original GrowCut, thus requires numerous amount of manual seeds, shown in the third column of Figure 4.3. On the contrary, the machine learned classifier is able to distinguish the dura on the brain surface from tumors by the use of the local shape features and the collection of focused training samples, and is able to provide automatic seeds in this dura area. Thus the boosting GrowCut reduces user manual seeds.

For cases 3 and 4, boosting GC improves the accuracy upon original GrowCut with minimum seeds (2 seeds in this study) and converges to the accuracy of original GrowCut with approximately 20 seeds. The reason is that these cases either included necrosis, dura or an additional unconnected tumor piece. Those structures cannot be correctly segmented by the original GrowCut, whereas they are correctly segmented by boosting GrowCut. As shown in case 3(the first column) of Figure 4.3, the original GrowCut mistakenly segments the connected dura and

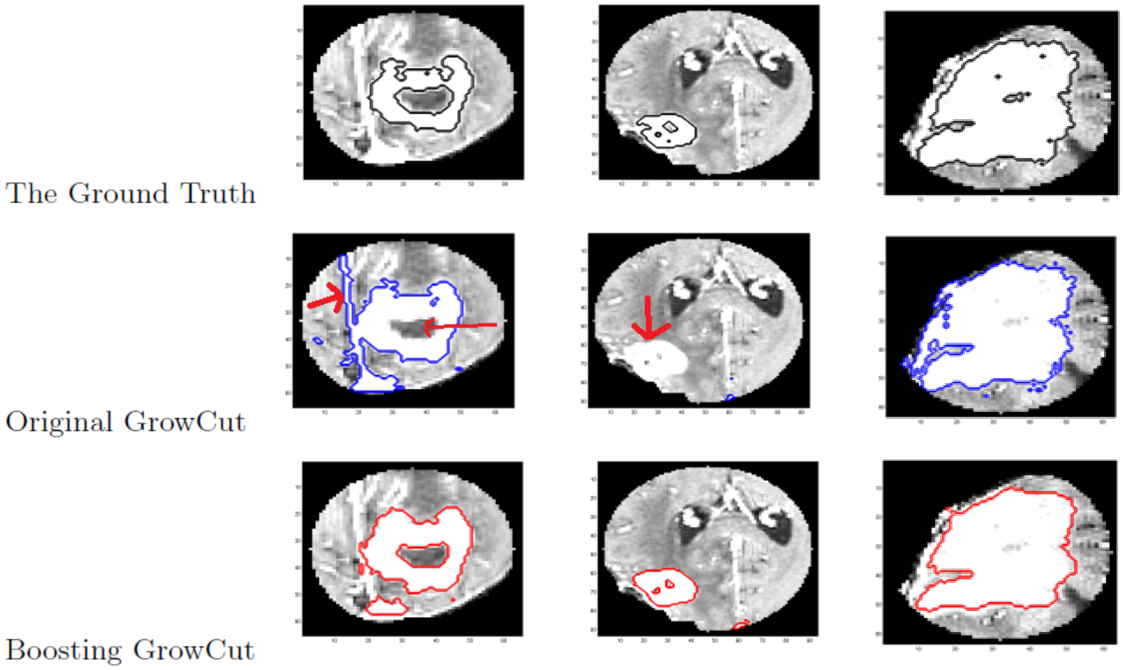


Figure 4.10: Examples of segmentation results. Columns: Case 3, 4 and 1; rows: the ground truth (black contour), the original GrowCut results with one tumor seed and one background seed (blue contour) and the boosting GrowCut results (red contour)

can not remove necrosis, while boosting GrowCut successfully overcomes both problems. In case 4 (the second column) of Figure 4.3, the tumor piece shown is a second unconnected component and is completely missed by the original GrowCut, whereas it is correctly segmented by boosting GrowCut.

For cases 5, 6 and 8 boosting GrowCut does not show obvious improvement. The main reason is that the image contains an isolated and contiguous tumor which does not include any of the structures mentioned above and original GrowCut generates good results as shown in Case 5 (the third column) of Figure 5.

For cases 7 and 10, boosting GrowCut improves the performance. In case 7, the contrast enhancement is fuzzy compared to the brain tissue, there is an unconnected component, and the tumor is attached to the brain surface dura. In case 10, the tumor has very irregular shape, the enhancement has thin connected pieces, and there are multiple necrosis. These problems can be handled by machine learning.

There are a few limitations in this study. First, we applied a linear classifier in the training phase to differentiate tumor class from the non-tumor class in the feature space. Future studies could apply a non-linear classifier to investigate whether a non-linear decision boundary improves performance. Additionally, multiple classifiers can be applied and compared to reduce the classifier bias. Second, to evaluate the interactive boosted GrowCut framework, we simulated additional seeds. In the future, it will be interesting to conduct a reader study and compare the amount of user interaction and the accuracy and consistency of the segmentation between original GrowCut and the proposed boosting GrowCut. Third, for automatic soft seeding, we applied an automatic detection of the dural region mask to limit the search region. In the future, we can investigate the use of brain atlas as prior information to replace the dural region detection.

The proposed framework shows potential in reducing the user interaction, improve accuracy and improve consistency in the preliminary dataset of 10 GBM tumor cases with different variations in tumor appearance. In the future, we will evaluate the method in a larger dataset of GBM tumor cases.

In conclusion, the preliminary results shows that fewer simulated seed points are needed for boosting GrowCut method to reach the same or better overlap ratio. It suggests that the accuracy may be improved and user interaction reduced. The error bars on the overlap ratio from multiple simulations are smaller for the boosted GrowCut method. It indicates that the result is more consistent when seed points are varied, and suggests that the reproducibility is improved. The framework for improving semi-automated segmentation using machine learning has the potential to be generalized to other applications.

4.5 Appendix: User Interaction Protocol to Collect Training Samples

- Scroll up and down of the tumor, find all tumors. Start from the biggest primary tumor.
  - If it is a gross tumor in the center slice, paint red in the middle of the tumor, find the farthest rim of the tumor, and paint blue in the background in the neighborhood of farthest rim.
  - If it is a ring shape in the center slice, red seeds have to be provided at 3 oclock, and 9 oclock directions. Blue seeds are still provided in the farthest rim of the tumor.
- For each tumor, go to the center slice of this tumor (or the slice with the biggest 2D diameter).
- A 3D sphere VOI will be generated.
- User has to scroll up and down to see whether the whole tumor is enclosed in the VOI.
  - If the missing piece is part of the gross tumor, user goes to the farthest slice of the missing piece, paint red on tumor parts and blue on neighboring non-tumor structures.
  - If the missing piece is a ring shape, go to the farthest slice and the farthest rim of the tumor, paint red on ring tumor and paint blue in both neighboring necrosis and brain tissues.
  - Repeat the last two steps till the primary brain tumor is completely enclosed.
  - If there is a 2nd, 3rd, 4th, tumor, if it is distant pattern, consider it as a separate tumor, and segment it later; if it is multi-focal, repeat the first two items till it is enclosed.

Distant: single new focus of enhancement or a qualitative assessment of recur-

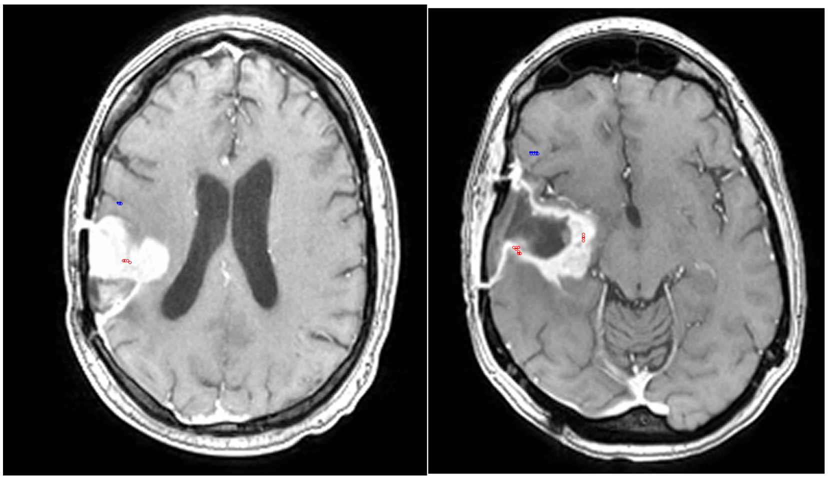


Figure 4.11: Two examples of user input seeds

rence centered more than 3 cm from the primary site resection cavity or margin of the primary residual tumor, which is mostly or all well defined.

Multifocal: more than one lesion site with each lesion having a mostly or completely well defined border with intervening areas of normal brain signal.

Figure 4.5 shows two examples of initial seeds suggested on the MR images with brain tumors.

CHAPTER 5

Between-scanner and Between-visit Variation in Normal White Matter Apparent Diffusion Coefficient Values in the Setting of a Multi-center Clinical Trial

Abstract

Purpose: To study the between-scanner variation and the between-visit reproducibility of brain ADC measurements in the setting of a multi-center chemotherapy clinical trial for GBM. Methods and Materials: ADC maps of 52 patients at six sites were calculated in-house from DW MR images obtained by seven individual scanner models of two vendors. The median and coefficient of variation (CV) of normal brain white matter (WM) ADC values from a defined region of interest (ROI) were used to evaluate the differences among scanner models, vendors, magnetic fields, as well as successive visits. Results: For baseline median ADC, no significant difference was observed between the different scanner models, different vendors, and different magnetic field strengths. For baseline ADC CV, a significant difference was found between different scanner models ($p = 0.0002$). No between-scanner difference was observed in ADC changes between two visits. For between-visit reproducibility, significant difference was seen between the ADC values measured at two successive visits for the whole patient group. Conclusion: The CVs varied significantly between scanners, presumably due to image noise. Consistent scanner parameter setup can improve reproducibility of

the ADC measurements between visits.

5.1 Introduction

Quantification of the reproducibility of measures based on diffusion weighted imaging is a prerequisite for the design of quantitative clinical studies using this modality. In the setting of a multi-center, multi-scanner chemotherapy clinical trial, it is necessary to evaluate the reproducibility of ADC measurements in order to reliably use them as a biomarker to evaluate GBM tumor treatment response.

The aims of our study were to evaluate between-scanner and between-visit variation of ADC measurement in the setting of a real-world multi-center drug trial where multiple centers and multiple scanners were involved. Even following radiation and variable chemotherapy, these patients still appear to have consistent normal brain WM ADC values by radiologist visual inspection. In our study, we evaluated quantitatively whether a variety of clinical MRI scanner models produce consistent measures of brain white matter ADC, and additionally, to evaluate the scan-rescan reproducibility in measured ADC at two visits between different scanner models. We examined not only absolute ADC values, but also the dispersion, as measure by the coefficient of variation (CV) of ADC measures. Data was obtained from a multi-center clinical trial for treatment of GBM.

5.1.1 Related Work

Different factors can introduce variations in ADC measurements, including patient age, number and strength of diffusion sensitizing gradients, field strength, location and size of the ROIs used for analysis, signal-to-noise ratio, and number of diffusion-sensitizing gradient directions. In a multi-center clinical trial, different scanner models may introduce variation too.

The effect of age is under debate. [EPP00] reported that advancing age is associated with a small but statistically significant increase of water diffusivity in brain white matter using 38 patients. However, another study [HSP02] using 80 healthy volunteers reported no difference among different age groups.

As for the number and strength of diffusion sensitizing gradients, [MIJ00] studied with one water phantom and 10 healthy volunteers scanned on one machine with $b = 0 - 800 s/mm^2$ using ROIs at different locations of the brain. They reported a significant decrease in the ADC values with increasing strength of diffusion sensitization, which can be explained, in part, by a more pronounced direct effect of microcirculatory perfusion on measures of water diffusivity at low b values. The study reported a significant difference when using a different number of b-values, but it may be due to the use of a low b-value. [BER98] compared a two-point and six-point calculation with 10 subjects using b-values between 0 and $1000 s/mm^2$, and reported a high correlation between the two techniques. [HSP02] did another study with high b-values (1000-3000), and found a high correlation between the two-point and multi-point methods, and little error in estimating ADCs calculated by the two-point method using high b-value. Steens et al examined different b-values and scan-rescan reproducibility on the whole-brain ADC histogram [SAS04].

[LFJ07] studied the effects of diffusion schemes and reported that the optimized PE6, PE10, PE15 and Jones30 schemes tested in their study have comparable precision. This means that they have comparable power to discriminate normal from abnormal. The observed differences in the DTI contrast due to different DW schemes are shown to be small relative to intra-session variability. This result suggests that typical clinical studies, which use similar protocols but different DW schemes, are readily comparable within the experimental precision. [HBM06] also concluded that the number of diffusion directions did not have a significant effect on reproducibility.

[HMS01] studied the influence of magnetic field strength and concluded that at a value greater than the SNR threshold of 20, there is no significant differences in mean diffusivity between 1.5 and 3.0 T. Their results suggest that as long as this threshold is observed, there is a negligible effect on mean diffusivity between diffusion- tensor MR imaging studies at 3.0 and 1.5 T. Inconsistent with [HMS01]'s study, [HLB06] reported that ADC values for gray and white matter were statistically significantly lower at 3.0 Tesla compared with 1.5 Tesla. Comparative clinical studies using ADC values should consequently compare ADC or FA results with normative ADC values that have been determined for the field strength used. The field-strength-related SNR gain they observed was in the range of 20% comparing 1.5 with 3.0 Tesla.

Location and size of the ROI may introduce the intra-reader or inter-reader variation of ADC measurement. [BU04] conducted a study involving two readers placing ROIs on eight different anatomical structures for 27 patients. They found that ADC values are found to be unreliable for assessing brain disease in some specified areas of the brain owing to interobserver variance and different ROI sizes [HBM06]. The reliability of ROI measurements has also been shown to vary regionally. The reproducibility worsens on the edge of structures and it is recommended that small-sized ROIs should preferably be drawn on areas of high anisotropy away from the edges.

The studies above are all single-scanner studies, however, it is essential to evaluate the multi-center reproducibility in a clinical trial. Sasaki et al. [SYW08] obtained DW images with nearly identical parameters at 1.5 and 3.0T from 12 healthy volunteers at seven institutions by using 10 magnetic resonance (MR) imagers provided by four different vendors. They demonstrated that absolute ADC values can substantially vary among different coil systems, imagers, vendors, and magnetic field strengths. One study [KBC09] reported that ADC measurements were highly reproducible in a two-center clinical trial and appear promising for evaluating the effects of drugs that target tumour vasculature.

5.2 Methods and Materials

5.2.1 Patient Group

A total of 68 patients with GBM brain tumors from six medical centers were obtained.

The patient selection criteria were: 1) no visible T2-weighted signal change in the white matter used for ROI analysis as determined by a neuro-radiologist; 2) no significant magnetic susceptibility artifacts and 3) normal brain structures clearly identifiable on the ADC map. Eleven patients did not satisfy the criteria. As a result, we had 57 patients with usable baseline scans.

The scanner selection criteria were: 1) scanners which scanned at least five patients, to increase substantial statistical power; 2) scanners which scanned the same patient at least 2 times (for the between-visit variation study).

As a result, we included 52 patients (31 men and 21 women; age range 19-78 years old; mean age, 52 years old) by 7 scanner models for between-scanner variation study, and 40 patients by 5 scanner models for the between-visit ADC reproducibility study. Patient age ranges for the seven scanners are: 1) 26-71; 2) 48-69; 3) 36-63; 4) 19-64; 5) 37-70; 6) 39-69; 7) 25-78. These patients were treated with radiation and chemotherapy, with normal appearance of brain white matter by visual assessment.

5.2.2 Scanner Protocol

Seven scanners from six centers with variability in scanner parameters provided a real-world ADC variation study. The seven scanners included two 3T scanners (Siemens TrioTim at two sites) and five 1.5T scanners (GE SIGNA HDx at two sites, GE SIGNA EXCITE, Siemens Avanto, Siemens Symphony). The protocol required use of a DW spin-echo (SE) echo-planar imaging (EPI) technique

	Scanner Model	Number of patients	Field Strength	DWI or DTI	Number of Diffusion Directions	In-plane Image Resolution	Slice Thickness
1	GE SIGNA HDx at site 1	8	1.5T	DWI or DTI	6 for DTI and N/A* for DWI	256*256	5mm
2	GE SIGNA HDx at site 2	7	1.5T	DWI	N/A*	256*256	5mm
3	GE SIGNA Excite	5	1.5T	DWI	N/A*	256*256	5mm
4	Siemens Symphony	5	1.5T	DTI	6	128*128 or 256*256	5mm
5	Siemens Avanto	9	1.5T	DTI	6	256*192 or 256*256	5mm
6	Siemens TrioTim at site 3	10	3T	DTI	30	128*128	5mm
7	Siemens TrioTim at site 4	8	3T	DWI	3	128*128	7mm

Table 5.1: Detailed protocols of the MR scanners

(TR=4000-12000ms, TE=60-110ms) using a b-factor between 700 and $1000s/mm^2$. 22-36 axial slices were acquired, with FOV = 220-240mm, slice gap 5-7mm, slice thickness 5-7mm, and in-plane image resolution 128*128, 256*192, or 256*256. The number of diffusion sensitization directions was from 3-30. The details of the scanner parameters are shown in Table 5.1. The data were anonymized and collected in the digital imaging and communications in medicine (DICOM) format.

5.2.3 Image Computation

All ADC maps were calculated from DW images with the same in-house software using a two-point method as shown in the following equation: $ADC = -\frac{lnS(b)-lnS(0)}{b}$ with b being the diffusion sensitivity factor ranging between 700 and $1000s/mm^2$, and S being the image intensity when $b = 0$ or $700s/mm^2$, and $b = 1000s/mm^2$. DW, trace DW or DT images were used to derive ADC values based on their availability. For DWI trace images, we calculated ADC maps from DW images by the equation above. For DTI, we calculate ADC for each orientation and average them to produce the final ADC map. Figure 5.1 shows two example ADC maps.

5.2.4 Study Design

A fixed size circular 2D ROI (radius = 7 pixels) was manually drawn on the normal-appearing brain white matter above the ventricles and confirmed by a

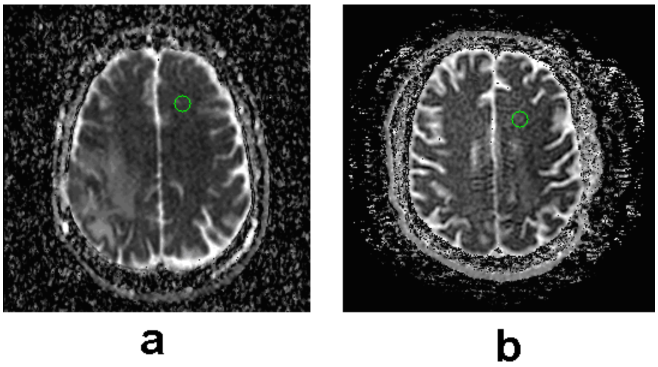


Figure 5.1: Two example ADC maps with brain WM ROIs. The window level and width are set to 2800 and $1500 * 10^{-6} mm^2/s$, respectively. The median and CV for each ROI is as follows: (A) median $= 721 * 10^{-6} mm^2/s$, CV = 0.11; (B) median $= 884 * 10^{-6} mm^2/s$, CV = 0.18.

board-certified neuro-radiologist. For each ROI, the ADC median and coefficient of variation (CV) were calculated. Median ADC value was used to represent the whole ROI ADC measurements to compare the absolute ADC value across different scanner models. CV, defined as the ratio of standard deviation (STD) to the mean, was used to evaluate the dispersion of the whole ROI ADC measurements. Median was used rather than mean to lessen the influence of random noise. CV was used instead of standard deviation (STD), because the STD of data should be understood in the context of the mean of the data. Figure 5.1 shows two images with different image quality, and thus different CVs and medians. Yasemin Bilgili et al. [BU04] reported that varying ROI sizes in brain WM did not yield statistically different ADC values.

Baseline ROI median and CV were used to explore the ADC variation across different scanners. Furthermore, the median change and CV change between two visits (typically 5-7 weeks apart) were calculated for each brain WM ROI, and differences across scanners were compared. What is more, between-visit reproducibility of ROI median and CV were evaluated for the whole patient group.

5.2.5 Statistical Analysis

Box-Cox transformation followed by the Shapiro-Wilk normality test were used to ensure the normality assumption in ADC median and CV were met. For baseline inter-scanner variation analysis, differences in baseline ROI ADC median and CV were examined by the ANOVA test among different scanners, different magnetic fields, and different scanner manufactures.

For the between-visit ADC change analysis, differences in the ROI ADC median change and CV change among different scanners were examined by the ANOVA test. Between-visit ADC measurement agreement was examined using an intra-class correlation coefficient (ICC), and difference in ADC between the two visits was evaluated by using a paired t-test.

5.3 Results

The Shapiro-Wilk normality test showed that the ROI ADC median ($p = 0.0014$) and CV ($p = 0.0065$) were not normally distributed. After log transformation of CV and inverse transformation of the median based on the Box-Cox model, data were normally distributed and thus eligible for the ANOVA test.

The results showed that there was no significant difference in median ADC ($p = 0.165$) between any two of the seven scanner models. However there was a significant difference in CV ($p = 0.0002$). Multiple comparison test by Tukey's HSD method was conducted. The result showed that the differences came from the following scanner pairs with significance levels: 1-7($p = 0.033$), 2-7($p = 0.002$), 3-7($p = 0.00007$), 5-7($p = 0.009$), 6-7($p = 0.023$), with the scanner index numbers corresponding to those in Table 5.1. Figure 5.2 displays the box plots of the seven patient groups.

For inter-vendor difference, we combined the three 1.5T GE scanners (No.

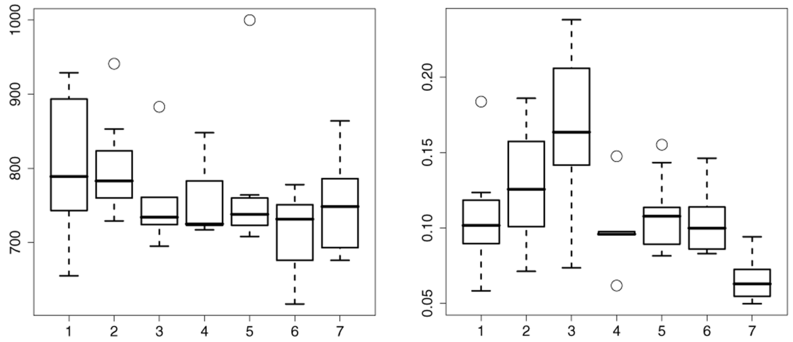


Figure 5.2: Box plots of the baseline brain WM ADC (units of $10^{-6}mm^2/s$) median and CV: (A) Brain WM ROI median; (B) Brain WM ROI CV.

1, 2, 3) patients into one group and the two 1.5T Siemens scanners (No. 4, 5) patients into another. We applied the ANOVA test to compare the difference in ROI median and CV between the two groups. There was no significant difference in either the median ($p = 0.30$) or CV ($p = 0.21$) for the two groups. Figure 5.3 displays the box plots of the aggregated groups.

We also evaluated the intra-vendor difference in ADC values when different magnetic field strength were used. We combined the two 1.5T Siemens scanners into one group and the two 3.0T Siemens scanners into another. The ANOVA test showed that there was no significant difference in terms of brain WM ROI median ($p = 0.16$). However, the test demonstrated a significant difference in brain WM ROI CV between different magnetic fields ($p = 0.04$). Figure 5.4 shows the box plots of the two groups.

For the between-visit reproducibility analysis, we had 39 patients by 5 scanners models with both baseline and follow-up data usable. The days between two visits are 34.35 ± 6.42, and the range is 27. For each patient, we calculated the ADC changes in ROI median and CV, and compared the difference across different scanners. The ANOVA test showed that there was no significant difference

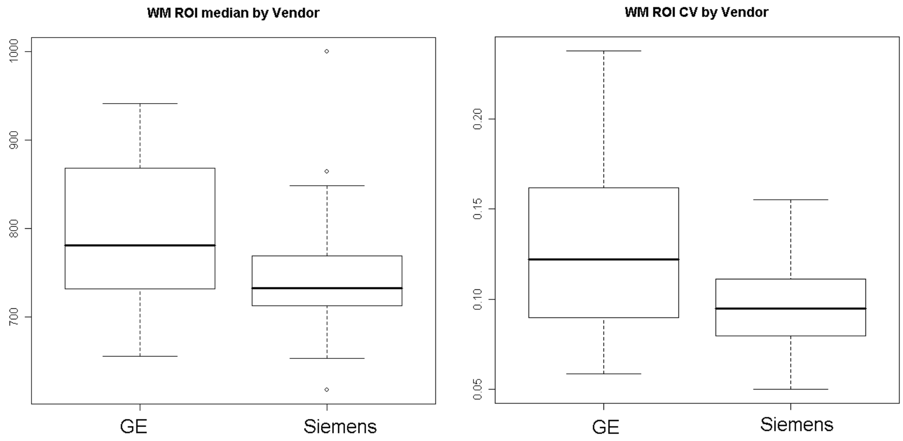


Figure 5.3: Box plots of the baseline brain WM ADC (units of $10^{-6}mm^2/s$) median and CV for all 1.5T GE and all 1.5T Siemens scanners: (A) Brain WM ROI median; (B) Brain WM ROI CV.

among the five scanners in median change ($p = 0.62$), and CV change ($p = 0.71$). Figure 5.5 displays the box plots of the five patient groups.

When the 39 patients were combined, a paired t-test showed that there was no significant difference between baseline and follow-up ADC values in CV ($p = 0.44$), but a significant difference in median ($p = 0.01$). The intra-class correlation coefficient (ICC) was 0.58 for the median, and 0.56 for the CV. Figure 5.6 shows the Bland-Altman plot of the ADC values measured at two time points before and after treatment. Assuming that normal WM is relatively unaffected by the treatment, the plot indicates that [-102.77, 150.51] is the range of normal ADC measurement variation.

Three of the patients from Scanner 1 did not have consistent acquisition parameters (number of diffusion sensitization directions) for baseline and follow-up scans. With these three patients excluded, the paired t-test between baseline and follow-up median ADC values showed that the significance level of $p = 0.05$ for the rest 36 patients. Figure 5.7 shows the box plot of the remaining 36 patients.

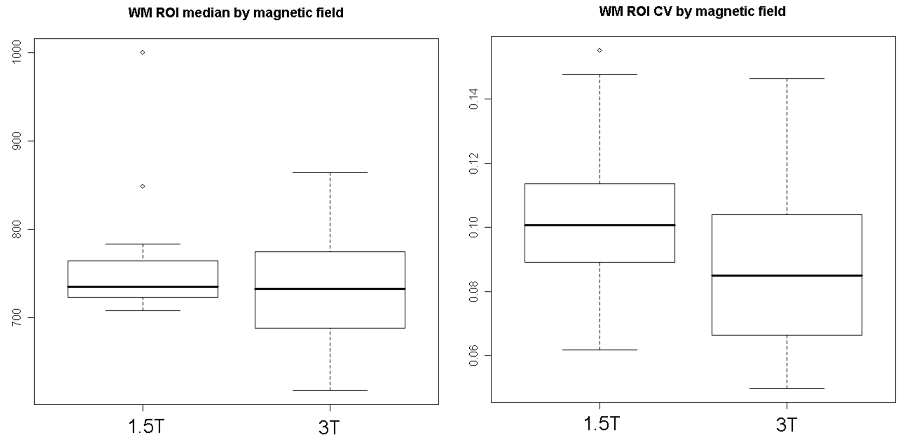


Figure 5.4: Box plots of the baseline brain WM ADC (units of $10^{-6} mm^2/s$) median and CV by magnetic field strength for all Siemens scanners: (A) Brain WM ROI median; (B) Brain WM ROI CV.

5.4 Discussion

In this study, we compared brain WM ADC measurements of patients with GBM tumors among different scanner models from different vendors at different medical sites with different field strengths and different acquisition styles. We acquired patient scans from a GBM drug clinical trial. The treatment of radiation and chemotherapy may affect the ADC values and cause them to change over time, reflecting real-world conditions in which patient treatment can be highly variable.

Others have found a small but significant difference in ADC values between scanners [SYW08]. We did not observe a significant difference in median ADC values across different scanner models. The power calculation showed that 52% power was achieved at the 0.05 level of alpha to detect differences in median ADC values among scanners, and that we needed 13 patients per scanner to have enough statistical power to detect the difference in ROI median. Therefore, we cannot conclude whether our observation arose because there was truly no difference, or because we did not have enough statistical power to detect a small change.

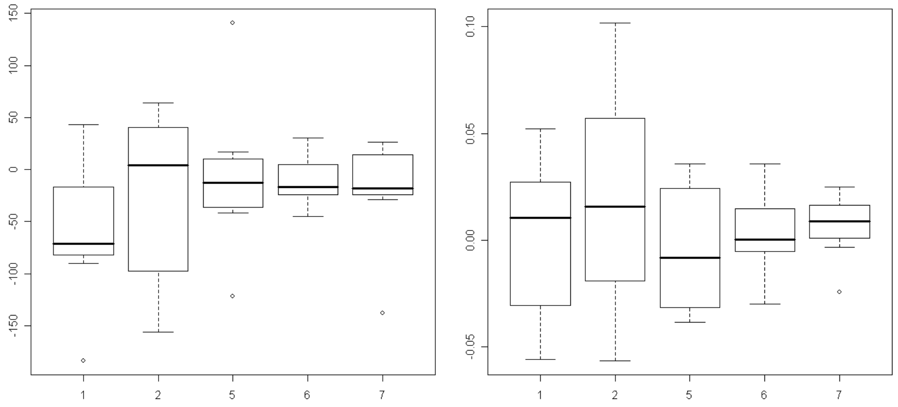


Figure 5.5: Box plots of the ADC changes for each patient between two visits: (A) Brain WM ROI median change by scanner; (B) Brain WM ROI CV change by scanner. The number of patients involved here are: Scanner 1: n=7; scanner 2: n=7; scanner 5: n=8; scanner 6: n=9; scanner 7: n=7. Scanners number corresponds to Table 5.1

In contrast to median ADC, we found a significant difference in CV of ADC measurements. This may be due to variability in image noise. Image noise can stem from many factors. First of all, different scanners from different vendors have different RF coil designs which can affect the accuracy of ADC values [SYW08]. Second, different magnetic field strength will lead to different signal to noise ratios, and thus different image quality (see below). Moreover, different acquisition techniques may result in different sensitivity. For instance, the DTI technique applies six or more gradient orientations to obtain the images, while the DWI technique uses three gradient orientations and obtains the averaged signal. Additionally, the total number of diffusion sensitization directions may also affect the accuracy of ADC measurement. In this study, there were both DWI and DTI scans with the number of diffusion sensitization directions varying from 3 to 30, as shown in Table 5.1. Moreover, different fields of view and slice thicknesses mean

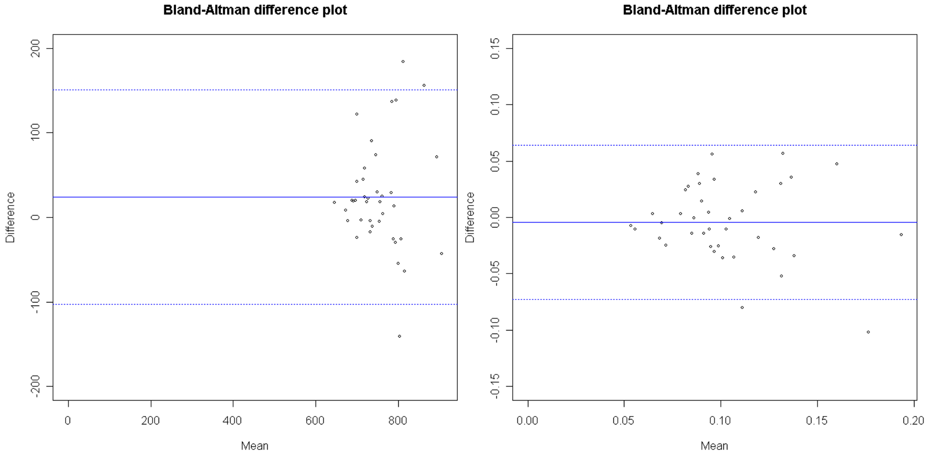


Figure 5.6: Bland-Altman plot to visualize the agreement between the ADC values measured at two time points before and after treatment: (A) median, Mean difference = 23.87, Upper agreement limit = 150.51, Lower agreement limit = -102.77; (B) CV, Mean difference = -0.004, Upper agreement limit = 0.06, Lower agreement limit = -0.07.

that a single voxel represents a different physical volume, resulting in a different signal-to-noise ratio. Lastly, image post processing, including image interpolation, filtering to improve image quality, etc may have variable effects.

For inter-manufacture variability, we compared the 1.5T Siemens scanners and 1.5T GE scanners and did not observe significant difference in median or CV of ADC measurements. Our observation agrees with the report by Koizumi et al [KMK03]. Koizumi et al also reported a good relationship in ADC values between scanners given a proper b factor in their phantom study [KMK03].

For ADC variability between different magnetic field strengths, we compared the 1.5T and 3T scanners from the same manufacture (Siemens). We observed no difference in median ADC values, consistent with prior studies [BU04, SYW08]. However, we observed that CV of ADC measurements from 1.5T Siemens scanners was larger than 3.0T Siemens scanners, which meant, ADC measurements

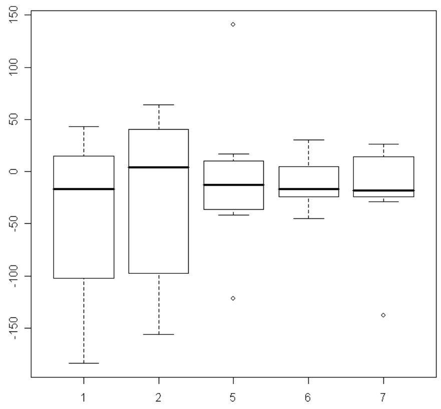


Figure 5.7: Box plots of the brain WM ROI median changes by scanner between two visits with 36 patients

from 1.5T scanners were more dispersed than 3.0T scanners. The result is logical since signal to noise ratio (SNR) increases with higher magnetic field strength, resulting in less noise and less dispersion. Interestingly, with our pooled data, the 1.5T Siemens scanners used a DTI acquisition technique and post-processed the images with interpolation, while the 3.0T Siemens scanners used a DWI acquisition technique without interpolating the raw images. Both of these factors may affect the dispersion of the brain WM ADC values. With our data, we were not able to evaluate their separate effects.

Besides analyzing between-scanner ADC variation with baseline data, we also explored the between-visit ADC variation among multiple scanners and ADC reproducibility between two successive visits. The interval between two visits ranged from 5-7 weeks. We found no significant difference across different scanner models in median change or CV change between successive visits.

As for the between-visit reproducibility, ADC measurements did not show high reproducibility for the whole 39 patients group. Three of the patients from scanner 1 did not have consistent acquisition parameters (number of diffusion sensitization directions) for baseline and follow-up scans. With these three patients excluded, the paired t-test between baseline and follow-up median ADC values showed that the level of significance was decreased from 0.01 to 0.05 for the remaining 36 patients. We conclude that consistent scanner parameters are necessary to achieve good between-visit reproducibility.

These data suggests that ADC measurements can have good reproducibility between two successive visits with exactly the same scanner parameters. The between-visit ADC change does not vary significantly among scanners.

The primary limitation of our study was that relatively few patients were available for each scanner model, diminishing the statistical power of our analysis, and thus the ability to detect small changes. For future studies, a larger subject population is required to increase statistical power to detect more pairwise differences. Moreover, we were not able to assess the degree of between-scanner ADC variation due to a lack of a controlled (phantom) study. Lastly, inconsistent between-scanner parameters (DW gradient orientations, number of b-factor, and in-plane image resolution) may introduce bias. However, in a real-world multicenter clinical trial, it is possible that scanner parameters are set differently among different medical sites.

In conclusion, we performed a comparison study in a real-world clinical trial to determine if ADC measurements were consistent across different scanners between different visits. For between-scanner ADC variation, the results showed a significant CV difference in ADC measurements across different scanners. Median difference of ADC measurements may be found given more patients and more statistical power. Moreover, CV difference was reported for different magnetic field strength and CV was smaller for 3T than 1.5T. For between-visit ADC variation, the ADC measurements can have good reproducibility with consistent scanner parameters between two successive visits 5-7 weeks apart. Furthermore, the ADC measurement changes did not vary significantly across scanners in terms of both median change and CV change. This implies that ADC changes before and after treatment have potential as a surrogate endpoint. For studies using baseline ADC as treatment predictors, we suggest evaluating image quality by use of brain WM.

CHAPTER 6

CADrx for GBM Brain Tumors: Predicting Treatment Response from Changes in Diffusion-Weighted MRI

This chapter is based on the manuscript "CADrx for GBM Brain Tumors: Predicting Treatment Response from Changes in Diffusion-Weighted MRI", by J. Huo, K. Okada, H.J. Kim, W.B. Pope, J.G. Goldin, J.R. Alger and M.S. Brown, in *Algorithm*, 2009, 2(4), 1350-1367.

Abstract

The goal of this study was to develop a computer-aided therapeutic response (CADrx) system for early prediction of drug treatment response for glioblastoma multiforme (GBM) brain tumors with diffusion weighted (DW) MR images. In conventional Macdonald assessment, tumor response is assessed nine weeks or more post-treatment. However, we will investigate the ability of DW-MRI to assess response earlier, at five weeks post treatment. The apparent diffusion coefficient (ADC) map, calculated from DW images, has been shown to reveal changes in the tumor's microenvironment preceding morphologic tumor changes. ADC values in treated brain tumors could theoretically both increase due to the cell kill (and thus reduced cell density) and decrease due to inhibition of edema. In this study, we investigated the effectiveness of features that quantify changes from pre- and post-treatment tumor ADC histograms to detect treatment response. There are three parts to this study: first, tumor regions were segmented on T1w contrast enhanced images by Otsu thresholding method, and mapped from T1w images onto ADC images by a 3D region of interest (ROI) mapping tool using DICOM header information; second, ADC histograms of the tumor region were extracted from both pre- and five weeks post-treatment scans, and fitted by a two-component Gaussian mixture model (GMM). The GMM features as well as standard histogram-based features were extracted. Finally, supervised machine learning techniques were applied for classification of responders or non-responders. The approach was evaluated with a dataset of 85 GBM patients under chemotherapy, in which 39 responded and 46 did not, based on tumor volume reduction. We compared adaBoost, random forest and support vector machine classification algorithms, using ten-fold cross validation, resulting in a best accuracy of 69.41% and corresponding area under the curve (Az) of 0.70.

6.1 Introduction

Computer aided diagnosis (CADx) can be defined as a diagnosis that is made by a radiologist who uses the output from a computerized analysis of medical images as a “second opinion” in detecting lesions and for making diagnostic decisions [Gig00]. One aim of a typical CADx system is to extract and analyze the characteristics of lesions in an objective manner to aid the radiologist. Here, the “diagnostic” decision relates to treatment response and early classification of drug responders versus non-responders, and we name our proposed system a computer-aided therapeutic response assessment (CADrx) system.

Glioblastoma multiforme (GBM) is the most aggressive and lethal primary brain tumor in humans. Anti-angiogenesis drugs are increasingly being explored in clinical trials as therapeutic options. In a phase II in vivo clinical trial, the conventional way to assess treatment response is tumor size change after chemotherapy or radiotherapy based on Macdonald criteria and evaluated on T1-weighted contrast enhanced (T1wCE) MR images. However, efficacy can only be evaluated at least 8-10 weeks after treatment.

Diffusion weighted magnetic resonance imaging (DW-MRI) has the potential as a surrogate biomarker to reveal changes in the tumor microenvironment that precede morphologic changes [RML03]. DW-MRI depends on the microscopic mobility of water. This mobility, classically called Brownian motion, is due to thermal agitation and is highly influenced by the cellular environment of water. Because water diffusion is strongly affected by molecular viscosity and membrane permeability between intra- and extracellular compartments, DW-MRI can be used to characterize highly cellular regions of tumors versus acellular regions. Treatment response can be manifested as a change in tumor cellularity, which may precede tumor size changes. Thus, findings on DW-MRI can be an early sign of biologic changes [PLM09].

The purpose of this study is to use the apparent diffusion coefficient (ADC), derived from DW-MR images, for early prediction of the tumor volume change on a later scan. Once GBM brain tumors are segmented on T1wCE images, the tumor ROI is mapped onto derived ADC maps and the histogram of tumor ADC values is extracted for automatic treatment response prediction.

Diffusion MRI has been explored for early detection of GBM brain tumor treatment response before the tumor size changes. Table 6.1 presents a review of the recent studies that used DWI for GBM early prediction of treatment response. Ross et al [RML03] reported a significant ADC value increase in effective therapeutic intervention in pre-clinical studies and presented two patients to support this hypothesis in a preliminary clinical study. Mardor et al. [MPS03] applied

Authors	Method	Number of patients
Chenevert et al. [CST00]	mean ADC	2
Ross et al. [RML03]	mean ADC	2
Mardor et al. [MPS03]	mean ADC, diffusion index	10
Moffat et al. [MCM06b]	functional diffusion map	20
Hamstra et al. [HCM05]	functional diffusion map	34

Table 6.1: Summary of related methods in GBM tumor treatment response using DWI

both low and high b-value and used mean ADC and diffusion index for treatment response evaluation. Moffat et al [MCM06b] calculated voxel-by-voxel tumor ADC value changes over time and displayed them as a functional diffusion map for correlation with clinical response. They reported that the number of voxels with increased ADC is related to treatment efficacy. Our previous work [HKP09] showed promising results for using ADC histogram analysis, and in this chapter we further explore a more sophisticated classifier and designed experiments to evaluate the two-component histogram modeling.

Machine learning and statistical pattern recognition have made contributions to the biomedical community because they can improve the sensitivity and/or specificity of detection and diagnosis of disease, while at the same time increasing objectivity of the decision-making process [Saj06]. The need for machine learning is perhaps greater than ever given the dramatic increase in medical data being collected, with new detection and diagnostic modalities being developed, as well as the complexity of the data types and importance of multimodal analysis. Machine learning can provide new tools for interpreting the high-dimensional and complex datasets with which the clinician is confronted [Saj06]. In our study, we explored three different classification methods: AdaBoost, random forest, and support vector machine.

The AdaBoost algorithm, introduced by Freund and Schapire [FS99], is an iterative algorithm that can boost weak classifiers into a strong classifier and improve the final accuracy. In each iteration, a feature is used within a weak classifier and the best feature is selected to minimize the average training error. Then, the weights on training samples are redistributed in such a way that the weight of accurately classified samples will be reduced while the weight of misclassified samples is raised. Therefore, AdaBoost focuses on the most difficult samples [DHS00]. The final classifier aggregates the selected weak classifier from each iteration, and the weight for each weak classifier depends on its error rate. However, AdaBoost can be sensitive to noise and may introduce overfitting.

Random forests (RF) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees [Bre01]. Each tree is built with sampling the training cases with replacement. For each node of the tree, a subset of the features are randomly chosen to calculate the best split at that node. Each tree is fully grown and not pruned. Breiman suggests the generalization error for forests converges to a limit as the number of trees in the forest becomes large [Bre01]. The error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost but are more robust with respect to noise.

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression [Vap82, Bis95]. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are ”pushed up against” the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. SVMs have been reported to work well for pharmaceutical data analysis [BTB02].

There are two main challenges in this work. One challenge is the two competing effects in ADC changes after treatment. In general, water movement inside cells is more restricted than outside. Thus, increased cell density tends to lower ADC values, whereas increased edema (more interstitial water) results in higher ADC values. Therefore, theoretically, ADC values in treated brain tumors could not only increase due to the cell kill (and thus reduced cell density), but also decrease due to inhibition of edema. None of the listed studies above have specified the separate effects. Addressing this issue, we applied a two-component model to fit the tumor ADC histogram [PKH09a]. The other challenge is that it is difficult to directly identify GBM brain tumors on ADC maps. We developed a semi-automated framework to achieve this goal.

There are several contributions in this work. First, we developed a computer-aided method to semi-automatically identify tumors on ADC maps. Second, we explored the changes of different statistical features of the whole tumor ADC histogram. Moreover, we applied two-component Gaussian mixture modeling to fit the tumor ADC histogram to characterize the two competing effects. Next, we used the earth mover's distance (EMD) to directly measure the distance between the pre- and post-treatment tumor ADC histograms. Finally, we introduced a machine learning technique to perform feature selection and classification to differentiate responders and non-responders.

6.2 Image Acquisition

6.2.1 Patient Cohort

A total of 85 patients with GBM treated by anti-angiogenesis drugs were included in our preliminary study from our research database. Images in this database were acquired as part of a multicenter GBM treatment trial. Tumors were diagnosed by board-certified radiologists as responders or non-responders to drugs based on the Macdonald criteria from follow-up scans (8-10 weeks after baseline). The Macdonald criteria define tumor response by use of tumor size change, steroids, and neurological function. There are four response categories: complete response (CR): disappearance of enhancing tumors, off steroids, and neurologically stable or improved. Partial response (PR): $> 50\%$ reduction in size of enhancing tumor, steroids stable or reduced, neurologically stable or improved. Progressive disease (PD): $> 25\%$ increase in size of enhancing tumor or any new tumor, or neurologically worse, and steroids stable or increased. Stable disease (SD): all other situations. In our study, we used tumor volume to evaluate tumor size. More than 50% increase in volume is considered to be PD based on neuro-radiologists recommendations [HMB99]. Since GBM is a rapidly progressing disease, we classified PD as non-responders and CR, PR and SD as responders. As a result, 39 were responders and 46 were non-responders. The DW-MRI scans were performed 5-7 weeks apart between baseline and follow-up scans.

The patients in this study were pooled from six medical centers scanned on 9 different scanner models (GE/Siemens) including both 1.5T and 3T scanners. The imaging protocol for T1wCE is 3D volume in the axial plane with flip angle-spoiled gradient echo sequence (FSPGR) or magnetization-prepared rapid gradient-echo (MP-RAGE) sequence, 1-5 mm slice thickness, 0.9375 mm by 0.9375 mm pixel size, and 256*256 in-plane resolution. The imaging protocol for the DW images is either DWI or DTI, $700-1000 s/mm^2$ for b-value, 3-30 for the number of diffusion

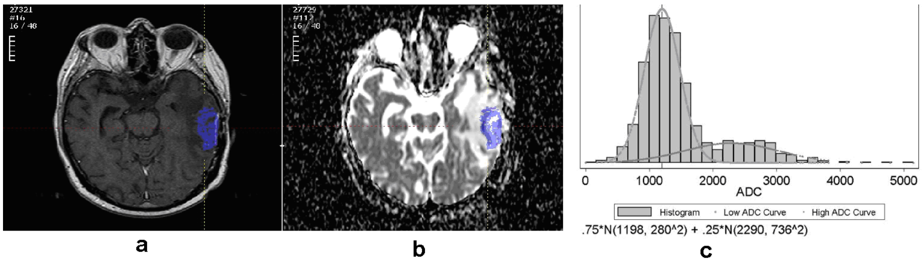


Figure 6.1: (a) An example of the tumor segmented on a T1wCE image; (b) An example of the tumor ROI mapped from T1wCE to ADC map; (c) An example of the tumor ADC histogram fitted by two-component Gaussian mixtures.

sensitization probing directions, 5-7 mm slice thickness, 1.797mm by 1.797 mm pixel size, and 256*256 or 128*128 in-plane resolution.

6.2.2 ADC Map Derivation

All ADC maps were calculated from DW-MR images with the same in-house software using a two-point method as shown in the following equation: $ADC = -\frac{ln[\frac{S(b)}{S(0)}]}{b}$, with b being the diffusion sensitivity factor ranging between 700 and $1000 s/mm^2$, $S(0)$ and $S(b)$ being the image intensity when $b = 0$ and $b = 700 - 1000 s/mm^2$. For DWI images, we calculated ADC maps from DW images by the previous equation. For DTI, we calculated ADC for each orientation and averaged them as the final ADC map. The calculation is the same as mentioned in Section 1.1.3. Figure 6.1(b) shows an example of a derived brain ADC map.

6.3 Semi-Automated Image Analysis on ADC Maps

All patients were scanned by both T1wCE MR images and DW-MR images. Since it is difficult to segment tumors accurately on derived ADC maps, we segmented tumors on T1wCE images first, and then mapped the 3D tumor contours onto the corresponding ADC maps.

6.3.1 Tumor Segmentation on T1wCE MR Images

All tumors were segmented on T1wCE images via a semi-automated method using the Otsu thresholding algorithm [Ots79] and seeded region growing [AB94]. First, radiologists drew a line from the inside of the tumor to the outside of the tumor on the approximate center slice of the tumor. Then intensity values along the line were collected to form a bimodal histogram, and the Ostu thresholding method was used to find the optimal thresholding value. Afterwards, 3D seeded region growing was applied to obtain refined segmentation results. Threshold-based segmentation methods are a standard approach to calculate tumor volume.

The Otsu thresholding method finds the threshold that minimizes the weighted within-class variation: $\sigma_w^2 = q_1(t)\sigma_1^2(t) + q_1(t)\sigma_2^2(t)$, with the class probability as $q_1(t) = \sum_{i=1}^{t} p(i)$, class variance as $\sigma_1^2(t) = \sum_{i=1}^{t}[i - \mu_1(t)]^2 \frac{p(i)}{q_1(t)}$ and class mean as $\mu_1(t) = \sum_{i=1}^{t} \frac{ip(i)}{q_1(t)}$. Given an initial $\mu_i(0)$ and $q_i(0)$, the algorithm does an exhaustive search by altering the thresholding value to find the optimal threshold.

Afterwards, seeded region growing [AB94] using the optimal threshold was applied to obtain the tumor contours in the 3D volume. Radiologists reviewed the results and made manual corrections when necessary. Figure 6.1(a) shows an example of a segmented tumor on a T1wCE image.

6.3.2 Tumor Mapping from T1wCE Images to ADC Maps

It is difficult for radiologists to directly delineate the tumor contours on ADC maps, and the scanner-provided T1w images and the derived ADC maps are not inherently co-registered, because they have different slice thickness, different field of view (FOV), and different image resolutions. Therefore, a 3D ROI mapping tool was developed to map the tumor ROIs from T1wCE images onto ADC maps based on the scanner geometry. Compared to the co-registration technique, the mapping tool only transformed voxels within the tumor ROI rather than the whole image volume; thus it was more computationally efficient. However, the mapping tool could not correct for patient motion; thus a board-certified radiologist was required to visually check the mapped contours and perform manual corrections when necessary.

The mapping tool used an affine transformation with the parameters extracted from the DICOM header based on physical locations. Equation 3 shows the calculation of the 3D physical location voxelwise. $\Delta_{i,j,k}$ is the physical voxel size read from the tag "pixel spacing" and calculated from "slice location"; $X_{x,y,z}$ and $Y_{x,y,z}$ are image orientation, read from the same tag "image orientation" which specifies the orientation of the image frame rows and columns, $Z_{x,y,z}$ is the z-direction orientation calculated from $X_{x,y,z}$, $Y_{x,y,z}$, $S_{x,y,z}$ is read from the tag "patient position" which specifies the physical location of the patients anterior-left-upper corner; i, j, k are voxel index; and $P_{x,y,z}$ are the calculated physical location of the voxel in millimeters. The transformation matrices are calculated for both the source and target ROIs respectively. For each voxel in the source ROI, the physical location is first calculated, and then the inverse operation is performed to calculate the corresponding voxel coordinates of the target ROI. Finally, radiologists visually check the contours on ADC maps and manually correct the tumor contours on ADC when necessary. Figure 6.1(b) shows an example of the mapped tumor ROI on the ADC map from the T1wCE image.

$$\begin{bmatrix} P_x \\ P_y \\ P_z \\ 1 \end{bmatrix} = \begin{pmatrix} X_x\Delta_i & Y_x\Delta_j & Z_x\Delta_k & S_x \\ X_y\Delta_i & Y_y\Delta_j & Z_y\Delta_k & S_y \\ X_z\Delta_i & Y_z\Delta_j & Z_z\Delta_k & S_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{bmatrix} i \\ j \\ k \\ 1 \end{bmatrix}$$

6.4 Feature Extraction and Classification

The differences between the features extracted from pre- and post-treatment tumor ADC histograms are used as the input to a tumor response classifier.

6.4.1 Observations

Figure 6.2 shows examples of tumor ADC histograms for both pre- and post-treatment from responders and non-responders. The upper histograms show the ADC distribution before the drug treatment, while the lower ones show the ADC distribution after the drug treatment. On the left is an example of a volumetrically responding tumor, while on the right is an example of a non-responding tumor. From the figure, we observe that not only the location but also the shape of the responder histogram changes after treatment. The two Gaussian mixture components change as well.

6.4.2 General Histogram Features

Different statistical features from tumor ADC histograms were extracted. According to clinical studies [CST00, RML03, MPS03, MCM06b, HCM05], ADC values should change after effective treatment. In our data set, we observed that the histograms exhibit changes not only in location, but also in shape. Therefore, we introduced the extraction of different ADC histogram features and explored changes in their pattern. The features are: mean, standard deviation, skewness, kurtosis, median, IQR (interquartile range), 25th percentile, and 75th percentile.

6.4.3 Features from GMM

Two-component Gaussian mixture modeling was applied to each tumor ADC histogram and the respective features were extracted. Due to the competing effects of

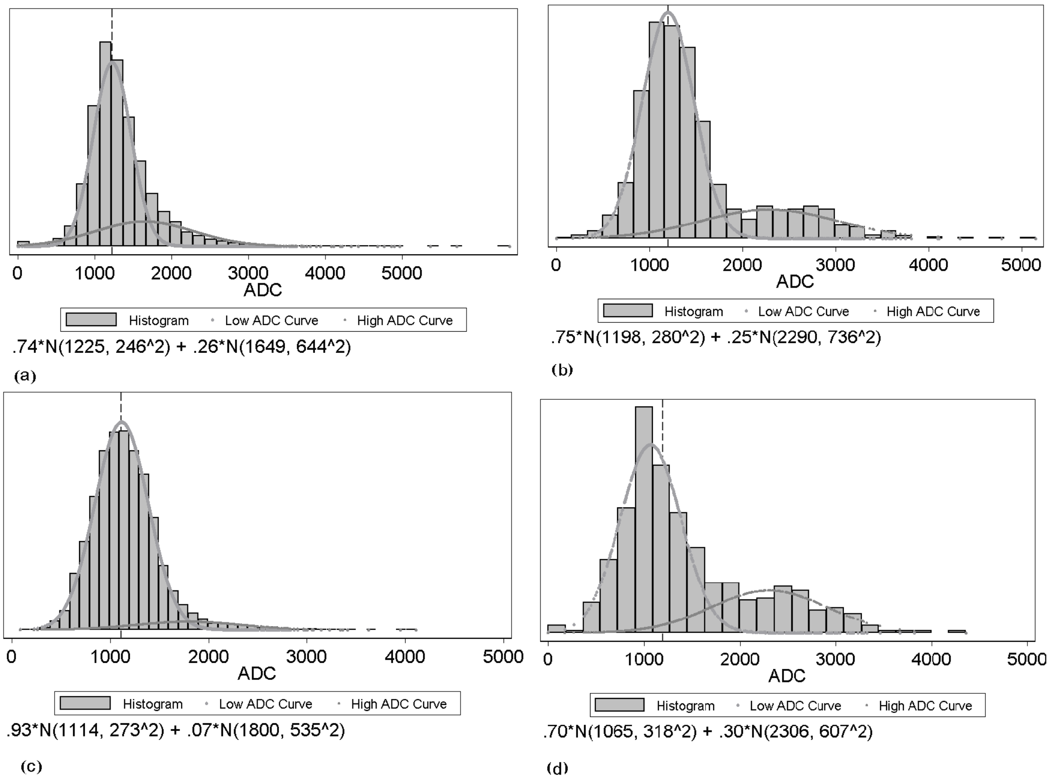


Figure 6.2: Examples of histograms from two tumors and two time points: (a),(c): example of a responding tumor for pre- and post-treatment respectively; (b),(d): example of a non-responding tumor for pre- and post-treatment respectively.

tumor cell density and edema, we made the assumption that the obtained tumor ADC histogram was composed of two components relating to tumor cellularity and edema. We assumed that the component with lower peak is influenced by tumor cellularity, and the component with higher peak by edema effects. We used a two component GMM to fit the ADC histogram for both baseline and follow-up scans and applied the EM algorithm to estimate GMM parameters, with x as the intensity values, α_i as the weight of the components, μ_i and σ_i as the Gaussian parameters.

$$f(x) = \sum_{i=1}^{2} \alpha_i G_i, G_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

The EM algorithm can be used to estimate the parameters of a parametric mixture model distribution: the weight of the components α_i, the Gaussian parameters μ_i, and σ_i. It is an iterative algorithm with two alternating steps: an expectation step (E-step) and a maximization step (M-step) [DLR77].

In the E-step, with the current parameter estimates of the mixture components, the algorithm calculates the expected values for the membership variables of all data points. At the $m+1$ iteration, the expectation is:

$$p_{ni}^{m+1} = \frac{\alpha_i^m G_i^m}{\sum_{i=1}^{2} \alpha_i^m G_i^m}$$

In the M-step, the algorithm maximizes the expectation value and updates the corresponding parameters. The following solutions can be developed:

$$\mu_i^{m+1} = \frac{\sum_{n=1}^{N} x_n p_{ni}^{m+1}}{\sum_{n=1}^{N} p_{ni}^{m+1}}, (\sigma_i^{m+1})^2 = \frac{\sum_{n=1}^{N} (x_n - \mu_i^{m+1})^2 p_{ni}^{m+1}}{\sum_{n=1}^{N} p_{ni}^{m+1}}, \alpha_i = \sum_{n=1}^{N} p_{ni}^{m+1}$$

The features we obtained from the GMM-EM were named as lower peak mean (LPM), lower peak variance (LPV), lower peak proportion (LPP), higher peak mean (HPM), higher peak variance (HPV) and higher peak proportion (HPP). Figure 6.2 shows examples of tumor ADC histograms fitted by GMM with low ADC and high ADC curves overlaid.

Combining GMM features with the statistical features, we obtained 14-dimensional feature vectors for both pre- and post-treatment tumor histograms. Then, we calculated the rate of change between the pre- and the post-treatment tumor histogram. Therefore, we computed a 14-dimensional difference feature vector.

6.4.4 Earth Mover's Distance

We applied the earth mover's distance (EMD) [RTG98, LO07] as a metric to directly evaluate the distance between the pre- and post-treatment tumor ADC histograms. Informally, if the histograms are interpreted as two different ways of piling up a given amount of dirt over the region D, the EMD is the minimum cost of turning one pile into the other; where the cost is the amount of dirt moved times the distance by which it is moved. The calculated EMD value was appended as the 15th element in the difference feature vector. The resulting 15-dimensional vector is used as the classification input for making our diagnostic decision.

6.4.5 Classification

We compared three representative classification techniques with different characteristics: AdaBoost, random forests (RF) and support vector machine (SVM) (as described in Section 3.2.1.3). We employed three classifiers to avoid biasing the results by selecting a single classification method. The reason we chose them is that the first two classifiers both include a feature selection mechanism. By applying these two classification techniques, we are seeking the best features that separate responders from non-responders. Moreover, SVM is reported to outperform several of the most frequently used machine learning techniques in structure activity relationship (SAR) analysis [BTB02]. In this study, all classifiers were implemented by using the open source data mining software Weka [WF05]. Their performance was evaluated using 10-fold cross validation.

Three analyses were performed. First, the conventional method of using mean ADC for treatment response classification was applied [RML03]. Second, the difference feature vectors of general statistical histogram features without GMM features were used. The AdaBoost, RF classifier, and SVM were applied, and results from the three classifiers were compared. Finally, all statistical features including the GMM features were used. The three classifiers were applied, and the results were compared. The results of accuracies from different classification techniques were compared with conventional method of ADC mean changes by the test of proportions.

Overlap Ratio	100%	95-100%	90-95%	80-90%	60-80%	0-60%
Number of patients	10	7	3	2	5	4

Table 6.2: Distribution of overlap ratios

6.5 Results

6.5.1 Segmentation Performance

Figure 6.3 shows four examples of segmentation on T1wCE images and the mapped results on the derived ADC maps.

For quantitative evaluation of the tumor segmentation mapping results, we randomly selected 31 subjects' baseline data. The 31 tumors are from an ADC mapping database, 20 of which have different image resolutions between the T1wCE and ADC images in all three dimensions and 11 of which have exactly the same 3D image resolution in both modalities. We calculated the overlap ratio between the mapped ROI generated automatically by the tool and an ROI corrected by a neuro-radiologist. The overlap ratio (OR) is defined as follows, where A and B are two tumor ROIs and size(.) is the number of voxels in that ROI.

$$2 * size(A \cap B)/(size(A) + size(B))$$

The results are shown in Table 6.2 with 20 out of 31 ROIs (64.5%) having an overlap ratio over 90%.

6.5.2 Classification Performance

Using the conventional method of mean ADC change (subjects with a mean ADC increase classified as responders and those with an ADC decrease as non-responders) [RML03, CST00], the accuracy is 29.4% (25/85), with a sensitivity of 17.95% and a specificity of 60.87% (see Table 6.3).

The experiment with AdaBoost involved 10 learning iterations. The RF classi-

Classifier	Sensitivity	Specificity	Accuracy	Az
Mean ADC change	17.95%	60.87%	29.4%	0.33

Table 6.3: Performance of conventional way

Classifier	Sensitivity	Specificity	Accuracy	Az
AdaBoost	45.45%	75%	63.53%	0.61
Random Forest	54.55%	73%	65.88%	0.66
SVM	27.27%	92.3%	67.06%	0.60

Table 6.4: Performance comparison among three classifiers without GMM features

fier was composed of 10 trees, each of which is constructed considering five random features. The SVM classifier used non-linear polynomial kernels and normalized all features.

The results for the experiment using only the general histogram features without GMM are shown in Table 6.4 with sensitivity, specificity, accuracy and area under the ROC curve (Az). The accuracies of the three classifiers are significantly different ($p < .0001$) comparing with accuracy of Table 6.3. The ROC curves are shown in Figure 6.4. The curve using conventional mean ADC was plotted by varying the threshold of the mean ADC change used for the classification, while the curve using the three ML techniques were plotted by Weka. Weka plots the ROC curves by varying the threshold on the probability assigned to the positive class.

With GMM features added, the three classifiers with the same parameter setups were applied to the data. The results are shown in Table 6.5 with sensitivity, specificity, accuracy and area under the curve (Az) of the ROC curve. The accuracies of the three classifiers are not significantly different ($p > .0001$) comparing with accuracy of Table 6.4. The ROC curves are shown in Figure 6.5.

Classifier	Sensitivity	Specificity	Accuracy	Az
AdaBoost	39.39%	80.77%	64.7%	0.60
Random Forest	51.52%	80.77%	69.41%	0.70
SVM	27.27%	92.3%	67.06%	0.60

Table 6.5: Performance comparison among three classifiers with GMM features

6.6 Discussion

Compared to using only the mean ADC value, the quantitative statistical histogram features and the proposed classification system tremendously improved the accuracy from 29.4% to 69.41% (Az increased from 0.33 to 0.70). The statistical analysis indicates that all three classifiers are significantly different from the conventional mean ADC method with our dataset. Compared to general statistical histogram features, the classification with GMM features using the random forest technique slightly improved the accuracy from 65.88% to 69.41%, while adaBoost and RF classifiers generated the same accuracy no matter whether GMM features were included. There is no significant difference between the three machine-learned classifiers.

The conventional mean ADC method performs worse than a random classifier (Az < 0.5). The reason is that conventionally researchers hypothesized that mean ADC increases because the tumor cell density decrease after an effective treatment. This assumption may not be valid for our dataset, because it involves in an anti-angiogenesis drug, which suppresses the cancer cell growth without necessarily killing tumor cells (decreasing their density) at an early stage (5-7 weeks). Another possible reason is that in our dataset many of the GBM tumors are recurrent and necrotic. The treatment tends to reduce necrosis and edema, which will diminish ADC. Essentially there are two competing processes at work: cell density, edema and necrosis [PKH09a].

Another recent study included features that capture spatial information in tumor heterogeneity features. Functional diffusion maps (fDM) [HCM05, MCM06b] are a popular technique using the ADC value increase or decrease voxel-by-voxel. Moffat et al. applied fDMs to 20 patients, classified patients into three categories (PR, SD and PD), and reported 100% accuracy [MCM06b]. However, the threshold they used for classification was determined from a single dataset of 20 patients used for both training and testing, while in our experiments, a cross validation analysis was performed. In Moffat et al's study, they explored the assessment of fractionated radiation therapy for different types of brain tumors with 20 patients scanned on the same scanner [MCM06b]. However, in our study, we focused on the GBM brain tumors treated by anti-angiogenesis drugs, which suppress the blood supply for the tumor cells and may not directly decrease the tumor cellularity. The difference in accuracy may come from the different treatment mechanism. Additionally, our dataset is from GBM drug trials across multiple sites, thus our preliminary study is an important contribution for exploring DWI as an early imaging biomarker in a real pharmaceutical drug trial. In future work, we will extract texture feature to include spatial information, and shape features will be extracted as well. By introducing richer feature set, we aim to include more information about tumors and further improve the performance of the classification system.

One limitation of this study is that we classified CR, PR and SD as responders for the ground truth to achieve a binary classification. Since SD and PR may have different patterns in terms of their ADC histogram change, a multi-category classification system will be explored in future work. Another limitation of the study is that we used the Macdonald criteria at the eighth or tenth week after treatment for determining treatment response. In future work, time-to-progression and survival time will provide a better endpoint to classify treatment response. Another limitation comes from the 3D ROI mapping tool. This tool is more computationally efficient compared to the co-registration techniques, but it cannot correct for patient motion. Therefore, in our study, a board-certified radiologist's visually checked and edited all segmentation results as needed. In the future, a more sophisticated registration method with an image similarity measure may improve the accuracy of the tumor contours on ADC maps, and consequently improve the accuracy of the extracted features and the classifier performance.

ADC values obtained on pre-operative MRI scans are reported to be of prognostic value in patients with glioblastoma [PKH09a]. The term "prognosis" refers to predicting the likely outcome of an illness. ADC, reported to be inversely proportional to tumor cellularity, is gaining interest in predicting GBM tumor prognosis. Our proposed framework now uses changes in DW-MRI as an early surrogate outcome biomarker; however, the framework with feature extraction and machine learning technique could be generalized to pre-treatment DW-MRI as a predictive biomarker.

In this study, we developed a CADrx framework with machine learning techniques to automatically predict tumor treatment response before size change using DW-MRI. In our preliminary study, our major contributions are extracting statistical ADC histogram features, applying GMM to model the ADC histogram to interpret the competing effects of cellular density and edema, and applying machine learning techniques using all the extracted features. Changes in cell density and edema may be reflected in ADC values before size changes are apparent on standard MRI sequences. Therefore, ADC holds promise as a biomarker, in determining both which tumors are more likely to respond to treatment and which tumors are actually responding.

In conclusion, this work shows that a CADrx system using quantitative ADC histogram features and a machine-learned classifier has better performance in treatment response assessment over conventional analysis using only a mean ADC value. This will have major implications for clinical trials.This work has potential clinical significance for early treatment response assessment in GBM.

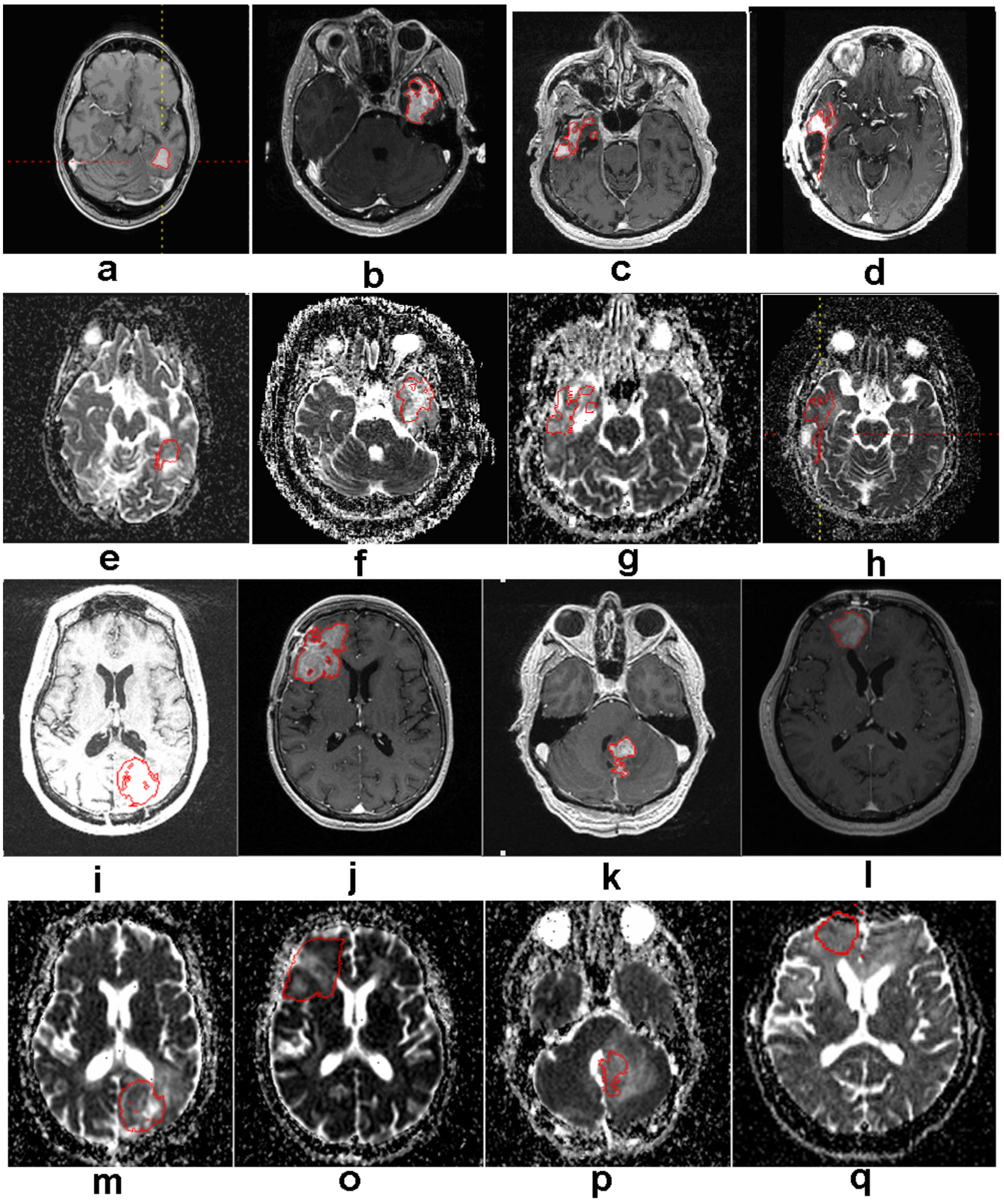


Figure 6.3: (a)-(d) and (i)-(l) show four examples of tumor segmentations on T1wCE images; (e)-(h) and (m)-(p) show the corresponding mapped tumor contours on ADC maps.

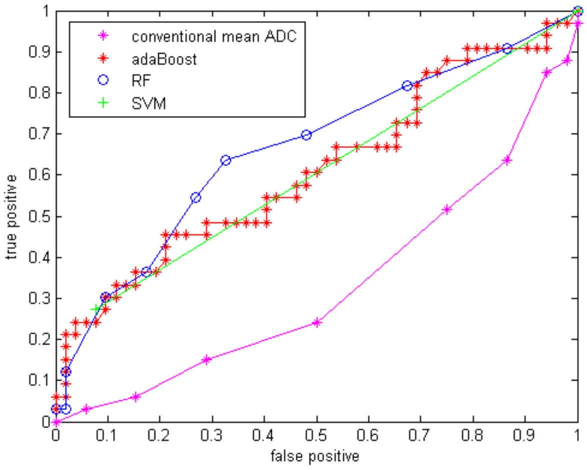


Figure 6.4: ROC curve for three classifiers without GMM features

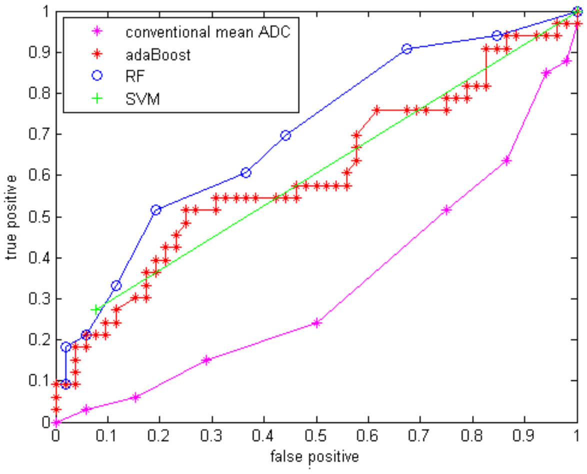


Figure 6.5: ROC curve for three classifiers with GMM features

References

[AA09] R Ayachi and NB Amor. "Brain Tumor Segmentation Using Support Vector Machines." In *ECSQARU '09: Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 736–747, Berlin, Heidelberg, 2009. Springer-Verlag.

[AB94] R Adams and L Bischof. "Seeded region growing." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **16**(6):641–647, 1994.

[AHH09] P Aljabar, RA Heckemann, A Hammers, JV Hajnal, and D Rueckert. "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy." *NeuroImage*, **46**:726–738, 2009.

[AM08] MM Ahmed and DB Mohamad. "Segmentation of Brain MR Images for Tumor Extraction by Combining Kmeans Clustering and Perona-Malik Anisotropic Diffusion Model." *International Journal of Image Processing*, **2**(1):2734, 2008.

[AMN10] Q Ain, I Mehmood, S Naqi, and M Jaffar. "Bayesian Classification Using DCT Features for Brain Tumor Detection." In Rossitza Setchi, Ivan Jordanov, Robert Howlett, and Lakhmi Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6276 of *Lecture Notes in Computer Science*, pp. 340–349. Springer Berlin / Heidelberg, 2010.

[BER98] JH Burdette, AD Elster, and PE Ricci. "Calculation of apparent diffusion coefficients (ADCs) in brain using two-point and six-point methods." *J Comput Assist Tomogr*, **22**(5):792–4, 1998.

[BHC93] JC Bezdek, LO Hall, and LP Clarke. "Review of MR image segmentation techniques using pattern recognition." *Medical Physics*, **20**:1033–1048, 1993.

[Bis95] CM Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Inc., New York, NY, USA, 1995.

[Bre01] L Breiman. "Random Decision Forest." *Machine Learning*, **45**:5–32, 2001.

[BS10] Z Beevi and M Sathik. "A Robust Segmentation Approach for Noisy Medical Images Using Fuzzy Clustering With Spatial Probability." *European Journal of Scientific Research*, **41**(3):437–451, 2010.

[BTB02] R Burbidge, MWB Trotter, BF Buxton, and SB Holden. “Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis.” *Computers and Chemistry*, pp. 5–14, 2002.

[BU04] Y Bilgili and B Unal. “Effect of region of interest on interobserver variance in apparent diffusion coefficient measures.” *AJNR Am J Neuroradiol*, **25**(1):108–11, 2004.

[CBS09] S Chandra, R Bhat, and H Singh. “A PSO based method for detection of brain tumors from MRI.” In *Nature Biologically Inspired Computing*, pp. 666–671, dec 2009.

[CDS07] W Chen, S Delaloye, DHS Silverman, C Geist, J Czernin, J Sayre, N Satyamurthy, WB Pope, A Lai, ME Phelps, and T Cloughesy. “Predicting Treatment Response of Malignant Gliomas to Bevacizumab and Irinotecan by Imaging Proliferation With [18F] Fluorothymidine Positron Emission Tomography: A Pilot Study.” *JOURNAL OF CLINICAL ONCOLOGY*, **25**:47144721, 2007.

[Cha06] S Cha. “Update on Brain Tumor Imaging: From Anatomy to Physiology.” *AJNR*, **27**:475–487, 2006.

[Cha11] MC Chamberlain. “Bevacizumab for the Treatment of Recurrent Glioblastoma.” *Clin Med Insights Oncol.*, **5**:117129, 2011.

[CHG98] MC Clark, LO Hall, DB Goldgof, R Velthuizen, R Murtagh, and MS Silbiger. “Automatic tumor segmentation using knowledge-based techniques.” *IEEE Trans. Med. Imaging*, **17**:187–201, 1998.

[CRD11] LC Hygino da Cruz Jr, I Rodriguez, RC Domingues, EL Gasparetto, and AG Sorensen. “Pseudoprogression and Pseudoresponse: Imaging Challenges in the Assessment of Posttreatment Glioma.” *Am J Neuroradiol*, **32**:1978–1985, 2011.

[CSD08] JJ Corso, E Sharon, S Dube, S El-Saden, U Sinha, and A Yuille. “Efficient Multilevel Brain Tumor Segmentation with Integrated Bayesian Model Classification.” *IEEE Trans. Med. Imaging*, **27**:629–640, 2008.

[CST00] TL Chenevert, LD Stegman, JMG Taylor, PL Robertson, HS Greenberg, A Rehemtulla, and BD Ross. “Diffusion Magnetic Resonance Imaging: an Early Surrogate Marker of Therapeutic Efficacy in Brain Tumors.” *JNCI J Natl Cancer Inst*, **92**(24):2029–2036, 2000.

[DCY08] S Dube, JJ Corso, A Yuille, TF Cloughesy, S El-Saden, and U Sinha. “Hierarchical Segmentation of Malignant Gliomas via Integrated Contextual Filter Response.” *Proc. SPIE 2008, 6914, 69143Y*, 2008.

[DHS00] RO Duda, PE Hart, and DH Stork. *Pattern Classification.* Wiley Interscience, 2000.

[DLR77] AP Dempster, NM Laird, and DB Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1):1–38, 1977.

[DRR10] S Drabycz, G Roldn, P de Robles, D Adler, JB McIntyre, AM Magliocco, JG Cairncross, and JR Mitchell. "An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging." *NeuroImage*, **49**:13981405, 2010.

[Dui00] RPW Duin. "PRTools - Version 3.0 - A Matlab Toolbox for Pattern Recognition." In *SPIE*, 2000.

[DVM09] R Drake, AW Vogl, and MitchellAWM. *Gray's Anatomy for Students.* 2009.

[EA66] RR Ernst and WA Anderson. "Application of Fourier Transform Spectroscopy to Magnetic Resonance." *Review of Scientific Instruments*, **37**:93–102, 1966.

[EMR11] BM Ellingson, MG Malkin, SD Rand, PS LaViolette, JM Connelly, WM Mueller, and KM Schmainda. "Volumetric analysis of functional diffusion maps is a predictive imaging biomarker for cytotoxic and anti-angiogenic treatments in malignant gliomas." *J Neurooncol.*, **102**:95103, 2011.

[EPP00] ST Engelter, JM Provenzale, JR Petrella, DM DeLong, and JR MacFall. "The effect of aging on the apparent diffusion coefficient of normal-appearing white matter." *AJR Am J Roentgenol.*, **175**(2):425–30, 2000.

[FHG01] LM Fletcher-Heath, LO Hall, DB Goldgof, and FR Murtagh. "Automatic segmentation of non-enhancing brain tumors in magnetic resonance images." *Artif Intell Med.*, **21**:43–63, 2001.

[FS99] Y Freund and RE Schapire. "A Short Introduction to Boosting." *Japonese Society for Artificial Intelligence*, 1999.

[Gig00] ML Giger. "Computer-aided diagnosis in medical imaging - A new era in image interpretation." *The world medical association medical imaging ultrasound*, pp. 75–78, 2000.

[Gra06] L Grady. "Random Walks for Image Segmentation." *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, **28**:11, 2006.

[GSI09] R Ghaemi, MN Sulaiman, H Ibrahim, and N Mustapha. "A Survey: Clustering Ensembles Techniques." *World Academy of Science, Engineering and Technology*, **50**, 2009.

[HBG02] S Ho, E Bullitt, and G Gerig. "Level set evolution with region competition: Automatic 3-d segmentation of brain tumors." In *Proceedings of International Conference on Pattern Recognition*, pp. 532–535, Quebec, Canada, August 2002.

[HBM06] E Heiervang, TE Behrens, CE Mackay, MD Robson, and H Johansen-Berg. "Between session reproducibility and between subject variability of diffusion MR and tractography measures." *Neuroimage*, **33**(3):867–77, 2006.

[HCM05] DA Hamstra, TL Chenevert, BA Moffat, TD Johnson, CR Meyer, SK Mukherji, DJ Quint, SS Gebarski, X Fan, CI Tsien, TS Lawrence, L Junck, A Rehemtulla, and BD Ross. "Evaluation of the functional diffusion map as an early biomarker of time-to-progression and overall survival in high-grade glioma." *Proc Natl Acad Sci USA*, **102**(46):16759–64, 2005.

[HHA06] RA Heckemann, JV Hajnal, P Aljabar, D Rueckert, and A Hammers. "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion." *NeuroImage*, **33**:115–126, 2006.

[HKP09] J Huo, HJ Kim, WB Pope, K Okada, JR Alger, Y Wang, JG Goldin, and MS Brown. "Histogram-based classification with Gaussian mixture modeling for GBM tumor treatment response using ADC map." *Proceedings of SPIE*, **7260**:72601Y–72601Y–7, 2009.

[HLB06] TA Huisman, T Loenneker, G Barta, ME Bellemann, J Hennig, JE Fischer, and KA Il'yasov. "Quantitative diffusion tensor MR imaging of the brain: field strength related variance of apparent diffusion coefficient (ADC) and fractional anisotropy (FA) scalars." *Eur Radiol*, **16**(8):1651–8, 2006.

[HMB99] SL Huhn, G Mohapatra, A Bollen, K Lamborn, MD Prados, and BG Feuerstein. "Chromosomal abnormalities in glioblastoma multiforme by comparative genomic hybridization: correlation with radiation treatment outcome." *Clin Cancer Res*, **5**(6):1435–43, 1999.

[HMS01] S Hunsche, ME Moseley, P Stoeter, and M Hedehus. "Diffusion-tensor MR imaging at 1.5 and 3.0 T: initial observations." *Radiology*, **221**(2):550–6, 2001.

[HRO11] J Huo, EM van Rikxoort, K Okada, HJ Kim, W Pope, J Goldin, and M Brown. "Confidence-based ensemble for GBM brain tumor segmentation." *Proc. SPIE Medical Imaging*, pp. 79622P–79622P–6, 2011.

[HSP02] J Helenius, L Soinne, J Perki, O Salonen, A Kangasmki, M Kaste, RA Carano, HJ Aronen, and T Tatlisumak. "Diffusion-weighted MR imaging in normal human brains in various age groups." *AJNR Am J Neuroradiol.*, **23**(2):194–9, 2002.

[HSS96] MA Hammoud, R Sawaya, W Shi, PF Thall, and NE Leeds. "Prognostic significance of preoperative MRI scans in glioblastoma multiforme." *J Neurooncol.*, **27**:65–73, 1996.

[HUG08] JW Henson, S Ulmer, and Harris GJ. "Brain Tumor Imaging in Clinical Trials." *AJNR Am J Neuroradiol*, **29**:419–424, 2008.

[KBC09] DM Koh, M Blackledge, DJ Collins, AR Padhani, T Wallace, B Wilton, NJ Taylor, JJ Stirling, R Sinha, P Walicke, MO Leach, I Judson, and P Nathan. "Reproducibility and changes in the apparent diffusion coefficients of solid tumours treated with combretastatin A4 phosphate and bevacizumab in a two-centre phase I clinical trial." *Eur Radiol*, **19**(11):2728–38, 2009.

[KCA09] H Khotanloua, O Colliotb, J Atifc, and I Bloch. "3D braintumor segmentationin MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models." *Fuzzy Sets and Systems*, **160**:14571473, 2009.

[KHD98] J Kittler, M Hatef, RPW Duin, and J Matas. "On Combining Classifiers." *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**(3):226–239, 1998.

[KMK03] K Koizumi, K Masuda, and M Komizu. "Report on the ECR2003 (European Congress of Radiology): Comparison of apparent diffusion coefficient (ADC) between different MRI scanners." *Nippon Hoshasen Gijutsu Gakkai Zasshi*, **59**(7):825–6, 2003.

[KT10] A Kassner and RE Thornhill. "Texture analysis: a review of neurologic MR imaging applications." *AJNR Am J Neuroradiol.*, **31**:809–16, 2010.

[KWN01] MR Kaus, SK Warfield, A Nabavi, PM Black, FA Jolesz, and R Kikinis. "Automated segmentation of mr images of brain tumors." *Radiology*, **218**:586591, 2001.

[Las10] A Lashkari. "A Neural Network based Method for Brain Abnormality Detection in MR Images Using Gabor Wavelets." *International Journal of Computer Applications*, **4**(7):1–8, July 2010. Published By Foundation of Computer Science.

[Lau89] PC Lauterbur. “Image formation by induced local interactions. Examples employing nuclear magnetic resonance.” *Clin Orthop Relat Res.*, **244**:3–6, 1989.

[LBE10] MO Leach, KM Brindle, JL Evelhoch, JR Griffiths, MR Horsman, A Jackson, G Jayson, IR Judson, MV Knopp, RJ Maxwell, D McIntyre, AR Padhani, P Price, R Rathbone, G Rustin, PS Tofts, GM Tozer, W Vennart, JC Waterton, SR Williams, and Workman P. “Assessment of antiangiogenic and antivascular therapeutics using MRI: recommendations for appropriate methodology for clinical trials.” *Br J Radiol.*, **9**:7–12, 2010.

[LFJ07] BA Landman, JA Farrell, CK Jones, SA Smith, JL Prince, and S Mori. “Effects of diffusion weighting schemes on the reproducibility of DTI-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5T.” *Neuroimage*, **36**(4):1123–38, 2007.

[LHK03] D LIPSITZ, RJ HIGGINS, GD KORTZ, PJ DICKINSON, AW BOLLEN, DK NAYDAN, and RA LECOUTEUR. “Glioblastoma Multiforme: Clinical Findings, Magnetic Resonance Imaging, and Pathology in Five Dogs.” *Veterinary Pathology*, **40**:659669, 2003.

[LO07] H Ling and K Okada. “An Efficient Earth Mover’s Distance Algorithm for Robust Histogram Comparison.” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **29(5)**:840–863, 2007.

[LSM05] CH Lee, M Schmidt, A Murtha, A Bistritz, J Sander, and R Greiner. “Segmenting brain tumor with conditional random fields and support vector machines.” In *in Proceedings of Workshop on Computer Vision for Biomedical Image Applications at International Conference on Computer Vision*, 2005.

[LUO05] J Liu, J Udupa, D Odhner, D Hackney, and G Moonis. “A system for brain tumor volume estimation via mr imaging and fuzzy connectedness.” *Comput. Med. Imaging Graph*, **29**:21–34, 2005.

[MBV02] N Moon, E Bullitt, K Van Leemput, and G Gerig. “Automatic brain and tumor segmentation.” *Proc MICCAI*, pp. 372–379, 2002.

[MCM06a] BA Moffat, TL Chenevert, CR Meyer, PE Mckeever, DE Hall, BA Hoff, TD Johnson, A Rehemtulla, and BD Ross. “The Functional Diffusion Map: An Imaging Biomarker for the Early Prediction of Cancer Treatment Outcome.” *Neoplasia*, **8**:259267, 2006.

[MCM06b] BA Moffat, TL Chenevert, CR Meyer, PE Mckeever, DE Hall, BA Hoff, TD Johnson, A Rehemtulla, and BD Ross. “The Functional Diffusion Map: An Imaging Biomarker for the Early Prediction of Cancer Treatment Outcome." *Neoplasia*, **8**(4):259267, 2006.

[MCS90] DR Macdonald, TL Cascino, SC Schold, and JG Cairncross. "Response criteria for phase II studies of supratentorial malignant glioma." *J Clin Oncol*, **8**(7):1277–80, 1990.

[MIJ00] ER Melhem, R Itoh, L Jones, and PB Barker. "Diffusion tensor MR imaging of the brain: effect of diffusion weighting on trace and anisotropy measurements." *AJNR Am J Neuroradiol.*, **21**(10):1813–20, 2000.

[MPS03] Y Mardor, R Pfeffer, R Spiegelmann, Y Roth, SE Maier, O Nissim, R Berger, A Glicksman, J Baram, A Orenstein, JS Cohen, and T Tichler. "Early detection of response to radiation therapy in patients with brain malignancies using conventional and high b-value diffusion-weighted magnetic resonance imaging." *J Clin Oncol*, **21**(6):1094–100, 2003.

[MSL04] MV Maldaun, D Suki, FF Lang, S Prabhu, W Shi, GN Fuller, DM Wildrick, and R Sawaya. "Cystic glioblastoma multiforme: survival outcomes in 22 cases." *J Neurooncol.*, **100**:61–7, 2004.

[OGA07] RA Ochs, JG Goldin, F Abtin, HJ Kim, K Brown, P Batra, D Roback, MF McNitt-Gray, and MS Brown. "Automated classification of lung bronchovascular anatomy in CT using AdaBoost." *Medical Image Analysis*, **11**(3):315–324, 2007.

[Ots79] N Otsu. "A threshold selection method from gray level histograms." *IEEE Trans. Systems, Man and Cybernetics*, **9**:62–66, 1979.

[PBH04] M Prastawa, E Bullitt, S Ho, and G Gerig. "A brain tumor segmentation framework based on outlier detection." *Medical Image Analysis Journal, Special issue on MICCAI*, **8**:275283, 2004.

[PBM03] M Prastawa, E Bullitt, N Moon, KV Leemput, and G Gerig. "Automatic brain tumor segmentation by subject specific modification of atlas priors." *Academic Radiology*, **10**:13411348, 2003.

[PCJ09] K Popuria, D Cobzasb, M Jagers, and SL Shaha. "3D variational brain tumor segmentation on a clustered feature set." In *SPIE medical imaging*, volume 7258, pp. 72591N–72591N–10, 2009.

[PID09] JM Provenzale, C Ison, and Delong D. "Bidimensional Measurements in Brain Tumors: Assessment of Interobserver Variability." *AJR Am J Roentgenol.*, **193**(6):W515–22, 2009.

[PKH09a] WB Pope, HJ Kim, J Huo, J Alger, MS Brown, D Gjertson, V Sai, JR Young, L Tekchandani, T Cloughesy, PS Mischel, A Lai, P Nghiemphu, S Rahmanuddin, and J Goldin. "Recurrent glioblastoma multiforme: ADC histogram analysis predicts response to bevacizumab treatment." *Radiology*, **252**(1):182–9, 2009.

[PKH09b] WB Pope, HJ Kim, J Huo, J Alger, MS Brown, D Gjertson, V Sai, JR Young, L Tekchandani, T Cloughesy, PS Mischel, A Lai, P Nghiemphu, S Rahmanuddin, and Goldin J. "Recurrent glioblastoma multiforme: ADC histogram analysis predicts response to bevacizumab treatment." *Radiology*, **252**:182–9, 2009.

[PLM09] AR Padhani, G Liu, D Mu-Koh, TL Chenevert, HC Thoeny, T Takahara, A Dzik-Jurasz, BD Ross, M Van Cauteren, D Collins, DA Hammoud, GJS Rustin, B Taouli, and PL Choyke. "Diffusion-Weighted Magnetic Resonance Imaging as a Cancer Biomarker: Consensus and Recommendations." *Neoplasia*, **11**(2):102125, 2009.

[PMB06] JM Provenzale, S Mukundan, and DP Barboriak. "Diffusion-weighted and perfusion MR imaging for brain tumor characterization and assessment of treatment response." *Radiology*, **239**:632–49, 2006.

[PSP05] WB Pope, J Sayre, A Perlina, JP Villablanca, PS Mischel, and TF Cloughesy. "MR imaging correlates of survival in patients with high-grade gliomas." *AJNR Am J Neuroradiol.*, **26**:2466–74, 2005.

[PVP95] WE Phillips, RP Velthuizen, S Phupanich, LO Hall, LP Clarke, and ML Silbiger. "Applications of fuzzy C-means segmentation technique for tissue differentiation in MR images of a hemorrhagic glioblastoma multiforme." *J. Magn. Reson. Imaging*, **13**:277–290, 1995.

[PXP11] WB Pope, Q Xia, VE Paton, A Das, J Hambleton, HJ Kim, J Huo, MS Brown, J Goldin, and T Cloughesy. "Patterns of progression in patients with recurrent glioblastoma treated with bevacizumab." *Neurology*, **76**(5):432–7, 2011.

[PYE11] WB Pope, JR Young, and BM Ellingson. "Advances in MRI Assessment of Gliomas and Response to Anti-VEGF Therapy." *Curr Neurol Neurosci Rep*, **11**:336344, 2011.

[Rij79] CJ van Rijsbergen. *Information Retrieval.* Butterworth, 1979.

[RML03] BD Ross, BA Moffat, TS Lawrence, SK Mukherji, SS Gebarski, DJ Quint, TD Johnson, L Junck, PL Robertson, KM Muraszko, Q Dong, CR Meyer, PH Bland, P McConville, H Geng, A Rehemtulla, and TL Chenevert. “Evaluation of cancer therapy using diffusion magnetic resonance imaging.” *Mol Cancer Ther*, **2**(6):581–7, 2003.

[RRM04] T Rohlfing, DB Russakoff, and CR Jr. Maurer. “Performance-Based Classifier Combination in Atlas-Based Image Segmentation Using Expectation-Maximization Parameter Estimation.” *IEEE TRANSACTIONS ON MEDICAL IMAGING*, **23**:8, 2004.

[RTG98] Y Rubner, C Tomasi, and LJ Guibas. “A Metric for Distributions with Applications to Image Databases.” *Proceedings of the 1998 IEEE International Conference on Computer Vision*, pp. 59–66, 1998.

[Saj06] P Sajda. “Machine learning for detection and diagnosis of disease.” *Annu Rev Biomed Eng*, **8**:537–65, 2006.

[SAS04] SC Steens, F Admiraal-Behloul, JA Schaap, FG Hoogenraad, CA Wheeler-Kingshott, S le Cessie, PS Tofts, and MA van Buchem. “Reproducibility of brain ADC histograms.” *Eur Radiol*, **14**(3):425–30, 2004.

[SJW04] SM Smith, M Jenkinson, MW Woolrich, CF Beckmann, TEJ Behrens, H Johansen-Berg, PR Bannister, M De Luca, I Drobnjak, DE Flitney, RK Niazy, J Saunders, J Vickers, Y Zhang, N De Stefano, JM Brady, and PM Matthews. “Advances in Functional and Structural MR Image Analysis and Implementation as FSL.” *Neuroimage*, **23**:208–219, 2004.

[SLG05] M Schmidt, I Levner, and R Greiner. “Segmenting Brain Tumors using Alignment-Based Features.” In *Fourth International Conference on Machine Learning and Applications*, 2005.

[SRH10] RN Sawlani, J Raizer, SW Horowitz, W Shin, SA Grimm, JP Chandler, R Levy, C Getch, and TJ Carroll. “Glioblastoma: a method for predicting response to antiangiogenic chemotherapy by using MR perfusion imaging–pilot study.” *Radiology*, **255**(2):622–8, 2010.

[SYW08] M Sasaki, K Yamada, Y Watanabe, M Matsui, M Ida, S Fujiwara, and E Shibata. “Variability in absolute apparent diffusion coefficient values across different platforms may be substantial: a multivendor, multi-institutional comparison study.” *Radiology*, **249**(2):624–30, 2008.

[SZE98] JG Sled, AP Zijdenbos, and AC Evans. “A nonparametric method for automatic correction of intensity nonuniformity in MRI data.” *IEEE Trans Med Imaging*, **17**:87–97, 1998.

[TOC10] S Taheri, SH Ong, and VFH Chong. "Level-set segmentation of brain tumors using a threshold-based speed function." *Image Vision Comput.*, **28**(1):26–37, 2010.

[Vap82] V Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

[VGK99] S Vinitski, CF Gonzalez, R Knobler, D Andrews, T Iwanaga, and M Curtis. "Fast tissue segmentation based on a 4d feature map in characterization of intracranial lesions Fast Tissue Segmentation Based on a 4D Feature Map in Characterization of Intracranial Lesions." *Journal of Magnetic Resonance Imaging*, **9**:768–776, 1999.

[VHN10] EG Van Meir, CG Hadjipanayis, AD Norden, HK Shu, PY Wen, and JJ Olson. "Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma." *CA Cancer J Clin.*, **60**(3):166–193, 2010.

[VK05] V Vezhnevets and V Konouchine. "GrowCut" - Interactive Multi-Label N-D Image Segmentation By Cellular.", 2005.

[VM11] H Veeraraghavan and JV Miller. "Active learning guided interactions for consistent image segmentation with reduced user interactions." *Biomedical Imaging: From Nano to Macro, IEEE International Symposium on*, pp. 1645–1648, 2011.

[VMV99] K Van Leemput, F Maes, D Vandermeulen, and P Suetens. "Automated model-based bias field correction of MR images of the brain." *Medical Imaging, IEEE Transactions on*, **18**(10):885 –896, oct. 1999.

[VUB03] MJ Vos, BMJ Uitdehaag, F Barkhof, JJ Heimans, HC Baayen, W Boogerd, JA Castelijns, PHM Elkhuizen, and TJ Postma. "Interobserver variability in the radiological assessment of response to chemotherapy in glioma." *Neurology*, **60**:826–830, 2003.

[Wal10] AD Waldman. "Magnetic Resonance Imaging of Brain Tumors - Time to Quantify." *Discovery medicine*, **9**:7–12, 2010.

[WF05] IH Witten and E Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2005.

[WKJ98] SK Warfield, MR Kaus, FA Jolesz, and Kikinis R. "Adaptive template moderated spatially varying statistical classification." *Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention*, 1998.

[WKM06] D Weishaupt, VD Koechli, and B Marincek. *How does MRI work? An Introduction to the Physics and Function of Magnetic Resonance Imaging.* 2006.

[WMR10] PY Wen, DR Macdonald, DA Reardon, TF Cloughesy, AG Sorensen, E Galanis, J Degroot, W Wick, MR Gilbert, AB Lassman, C Tsien, T Mikkelsen, ET Wong, MC Chamberlain, R Stupp, KR Lamborn, Vogelbaum MA, MJ van den Bent, and SM Chang. “Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group.” *J Clin Oncol*, **28**(11):1963–72, 2010.

[WRP08] P Wattuya, K Rothaus, JS Prani, and X Jiang. “A RandomWalker Based Approach to Combining Multiple Segmentations.” *ICPR*, 2008.

[WZW04] SK Warfield, KH Zou, and WM Wells. “Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation.” *IEEE Trans. Med. Imag*, **23**:903–921, 2004.

[ZME04] J Zhang, K Ma, MH Er, and V Chong. “Tumor Segmentation from Magnetic Resonance Imaging By Learning Via One-Class Support Vector Machine.” *International Workshop on Advanced Image Technology*, p. 207211, 2004.

[ZRL09] N Zhang, S Ruan, S Lebonvallet, Q Liao, and Y Zhu. “Multi-kernel SVM based classification for brain tumor segmentation of MRI multi-sequence.” In *ICIP’09: Proceedings of the 16th IEEE international conference on Image processing*, pp. 3337–3340, Piscataway, NJ, USA, 2009. IEEE Press.

[ZSS09] B Zhao, LH Schwartz, and Larson SM. “Imaging Surrogates of Tumor Response to Therapy: Anatomic and Functional Biomarkers.” *J Nucl Med.*, **50**(2):239–49, 2009.